

AD 697621
AGARD CONFERENCE PROCEEDINGS No. 39

AGARD CONFERENCE PROCEEDINGS No. 39

AGARD

ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT

7 RUE ANCELLE 92 NEUILLY SUR SEINE FRANCE

Storage and Retrieval of Information A User-Supplier Dialogue

★

JUNE 1968

DEC 9 1969

This document is available for public release and distribution

NORTH ATLANTIC TREATY ORGANIZATION

Reproduced by the
CLEARINGHOUSE
for Federal Scientific & Technical
Information Springfield Va. 22151



INITIAL DISTRIBUTION IS LIMITED
FOR ADDITIONAL COPIES SEE BACK COVER

200

NORTH ATLANTIC TREATY ORGANIZATION
ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT
(ORGANISATION DU TRAITE DE L' ATLANTIQUE NORD)

STORAGE AND RETRIEVAL OF INFORMATION
A USER-SUPPLIER DIALOGUE

Edited by

H. F. Vessey* and I. J. Gabelman†

* Chairman, Technical Information Panel, AGARD

† Avionics Panel, AGARD

1968

061.3"6.1968":
025.5:681.3.01



*Printed by Technical Editing and Reproduction Ltd
Harford House, 7-9 Charlotte St. London, W. 1*

FOREWORD

The increase in scientific and technical knowledge is creating a need for more efficient information storage and retrieval. These increases have in turn forced us to expand our search for more knowledge compounding the problem. Present manual and mechanical methods of storing and retrieving information are slow and cumbersome resulting in increasing interest in on-line interaction between computers and the data environment for information processing, retrieval and transfer.

Significant changes are beginning to occur in ideas on the way man uses information. Involved in these changes is the concept of the computer as an interactive medium. The concept of using a computer for routine clerical library jobs is being extended to include the adaptive, real-time, dynamic functions of modern computing machinery for storing and retrieving information. In experimental studies rapid memorizing and recall capabilities are being extended to vast quantities of information available in printed documents. With the prospect of automatic indexing and subsequent file search, a user may eventually be able to use these facilities to retrieve information from huge repositories, paralleling recall of information from his own memory. Effective harnessing of the computer power promises a revolution in the very nature of future libraries and information systems.

Current limitations in storage and retrieval are largely the result of limited memory capacity. One has to be satisfied with direct on-line inter-action with bibliographic information, followed by the time-honoured process of reading and manually sorting out the essential contents from many documents, a good number of which contain little, if any, really appropriate information. One can not yet expect to get specific answers directly to specific questions. Present day interactive bibliographic searches are effective in locating bulk substantive information, but are far from solving the basic problem of information transfer. This basic problem lies in the sorting out and assimilation of the desired specific information from the mass of documents that have been acquired.

Over a half dozen major international symposia have been held in the past two years on information storage, retrieval and dissemination. These have generally been slanted either to the documentalist or to the computer operator. Recognizing the need to fill in the gap between these two viewpoints, and the need for engineers and scientists in NATO member countries to widen their perspective on the subject of the processing and dissemination of technical information, the Technical Information and Avionics Panels of AGARD held a Symposium in Munich, June 18-20, 1968 on "Storage and Retrieval of Information - A User-Supplier Dialogue". The aim of the Symposium was to emphasize service to the real customer, the scientist or engineer.

This volume, the Proceedings of the Symposium, contains papers and discussions to stimulate and enlighten the scientists and engineers who are the actual, or potential, users of the systems described.

The subject is introduced by presenting the individual points of view of the user who asks for information and the supplier who stores information for subsequent retrieval. Present operational manual and mechanical systems are discussed to establish an appreciation of current practice. State-of-the-art in scientific and technical aids and the evolution of current methodology together with the user needs are applied to the development of potential future systems. Having an appreciation of the problems of storage and retrieval and possible solutions, the user-supplier loop is closed by discussing the dialogue which must exist between the two in order to obtain a successful operating system.

I. J. Gabelman
Advanced Studies Group,
Rome Air Development Centre, USA

and
H. F. Vessey
TIL Reports Centre,
Ministry of Technology, UK

AVANT-PROPOS

Le volume toujours croissant de nos connaissances scientifiques et techniques conduit à la nécessité de disposer de moyens plus efficaces d'emmagasinement et de sélection d'informations, volume qui amène à son tour à une recherche plus poussée de plus de connaissances, rendant encore plus complexe le problème. Les méthodes manuelles et mécaniques employées actuellement pour l'emmagasinement et la sélection des données étant lentes et lourdes, on s'intéresse de plus en plus à l'interaction directe entre les calculateurs et l'environnement des données en vue du traitement, de la sélection et du transfert d'informations.

Les idées concernant la façon dont les informations sont utilisées par l'homme commencent à se modifier de manière importante, et notamment, la notion du calculateur comme moyen d'interaction. La notion d'utiliser le calculateur pour assurer les tâches de nature systématique entreprises dans une bibliothèque se voit étendre aux fonctions adaptatives et dynamiques dans le temps réel des machines à calculer modernes destinées à l'emmagasinement et à la sélection des informations. Des études expérimentales permettent d'étendre à de vastes quantités d'informations consignées dans des documents imprimés les possibilités de mise en mémoire et de rappel rapides. Etant donné la perspective de moyens automatiques de classement et de recherche des dossiers, il est possible qu'un utilisateur puisse en définitive employer ces moyens pour sélectionner des informations dans de grands répertoires, de la même façon dont il se rappelle des informations de sa propre mémoire. Une utilisation efficace du pouvoir du calculateur fait penser aux bouleversements que pourrait subir la nature même des bibliothèques et des systèmes d'information de l'avenir.

Les limitations des systèmes actuels d'emmagasinement et de sélection de données sont dues en grande partie à une capacité de mémoire limitée. Il faut se contenter d'une interaction directe avec des informations bibliographiques, suivie du procédé consacré par l'usage que consiste à lire et à trier à main le contenu essentiel de bien des documents, dont un grand nombre comporte peu, s'il y en a de données vraiment utiles. On ne peut pas encore espérer obtenir des réponses spécifiques et immédiates à des questions spécifiques. Les recherches bibliographiques interactives que l'on fait actuellement permettent de localiser des renseignements "en vrac", mais sont loin d'être capables de résoudre le problème fondamental du transfert d'informations, problème qui réside dans le triage et l'assimilation des données particulières voulues, à partir de la vaste quantité de documents ayant été acquis.

Plus de six grands colloques internationaux ont été consacrés au cours de ces deux dernières années à la question de l'emmagasinement, de la sélection et de la diffusion d'informations. Ces colloques ont été en général destinés aux documentalistes ou aux opérateurs d'ordinateurs. Compte tenu de la nécessité de combler la lacune entre ces deux points de vue, et de la nécessité, pour les ingénieurs et les savants des pays membres de l'OTAN, d'élargir leur vue sur la question du traitement et de la diffusion des données techniques, les commissions "Informations Techniques" et "Avionique" de l'AGARD ont organisé un Symposium qui s'est tenu du 18 au 20 juin 1968 à Munich et qui a eu pour thème "L'Emmagasinement et la Sélection des Données - Dialogue entre Utilisateur et Fournisseur". Le colloque avait pour but de souligner les services à rendre au client réel, c'est-à-dire le savant ou l'ingénieur.

Le présent document, constitué par le Procès-Verbal de ce Symposium, comporte des exposés et des discussions destinés à encourager et à éclaircir les chercheurs et les ingénieurs qui sont les utilisateurs réels ou potentiels des systèmes y décrits.

Comme point de départ, on présente les points de vue individuels de l'utilisateur qui demande des informations et du fournisseur qui fait emmagasiner des informations en vue de leur sélection ultérieure. Les systèmes manuels ou mécaniques actuellement mis en oeuvre sont examinés pour établir une appréciation des usages courants. L'état actuel des connaissances dans le domaine des aides scientifiques et techniques, et l'évolution de la méthodologie actuelle, ainsi que les besoins de l'utilisateur, sont appliqués au développement de systèmes futurs possibles. Après avoir établi une appréciation des problèmes que posent l'emmagasinage et la sélection des données, et des solutions pouvant être trouvées, on fait la boucle utilisateur - fournisseur en évoquant le dialogue devant exister entre les deux si l'on veut réaliser un système opérationnel réussi.

I. J. Gabelman
Advanced Studies Group,
Rome Air Development Centre, USA

et

H. F. Vessey
TIL Reports Centre,
Ministry of Technology, UK

CONTENTS

	Page
FOREWORD	iii
AVANT PROPOS	iv
OPENING SPEECHES	
by Director Finn Lied Chairman of AGARD	viii
by Dr T. Benecke, AGARD National Delegate, Germany	ix
Paper 1 COMMUNICATION AND SECRECY IN SCIENCE by R. Schrader	1
Paper 2 THE SUPPLIER'S POINT OF VIEW - INTRODUCTORY PAPER by H.F. Vessey	11
Paper 3 LES PROBLEMES POSES PAR LE VOCASULAIRE DOCUMENTAIRE ET L'ORGANISATION DES DICTIONNAIRES ET THESAURUS par F. Levery (avec Appendice: LE CENTRE DE DOCUMENTATION DE LA CIE IBM FRANCE, CENTRE D'ETUDES ET RECHERCHES, LA GAUDE par R.J. Dubon)	21
Paper 4 FOUR "NEW" SCIENCES: AN APPROACH TO COMPLEXITY by E.B. Montgomery	41
Paper 5 TRENDS AND DEVELOPMENTS IN CHARACTER AND PATTERN RECOGNITION by L.A. Feidelman	49
Paper 6 EFFICIENT TRANSFER OF TEXTUAL INFORMATION by J.W. Altman	63
Paper 7 ON-LINE INFORMATION STORAGE AND RETRIEVAL by N.S. Prywes	77
Paper 8 NON-NUMERICAL MATHEMATICS AND DATA PROCESSING by F. Krückeberg	89
Paper 9 MANUAL SYSTEMS - TDCK CIRCULAR THESAURUS SYSTEM by J. A. Schüller	99
Paper 10 MECHANICAL SYSTEMS by N.E.C. Isotta	111
Paper 11 AN INTRODUCTION TO THE STUDY OF COST EFFECTIVENESS IN INFORMATION SYSTEMS by J.N. Wolfe	117
Paper 12 TECHNICAL INFORMATION SERVICES AND USER NEEDS by W.C. Christensen	123
Paper 13 SELECTIVE DISSEMINATION OF INFORMATION by M.S. Day	133

	Page
Paper 14 INTERACTIVE INFORMATION PROCESSING, RETRIEVAL, AND TRANSFER by J.C.R.Licklider	151
Paper 15 MAN-MACHINE INTERFACE by W.Händler	169
Paper 16 EDUCATION by F.Liebesny	179
 CONCLUSION	
SUMMING UP by R.Bree	187
VOTE OF THANKS by H.F.Vessey	192
NAME INDEX	193
SUBJECT INDEX	194

OPENING REMARKS

by Director Finn Lied
Director of AGARD

It gives me a great deal of pleasure to extend to you all a cordial welcome to this symposium on the storage and retrieval of information. I am very pleased to see that we have such a good representative of scientists and engineers, that is the users and producers of information, as well as documentalists and computer scientists. I should like also to welcome the considerable number of visitors and hope that they will find this introduction to the work of AGARD both profitable and enjoyable.

I wish to express our grateful thanks to the Federal German Government for its invitation to meet in this ancient city of Munich and for the ample facilities that have been provided so that you who are participating in the symposium will not only stimulate your professional interests in comfort but will also be able to enjoy the cultural and popular features of the historic capital of Bavaria and the magnificent scenery of the district.

The symposium programme is a joint effort by the two Panels on Avionics and Technical Information which are respectively responsible for the computer and documentation papers. With two completely different subjects such as these it is vitally important that there shall be full understanding between the two sets of practitioners so that what the one can supply agrees as nearly as possible with what the other needs.

There is, however, another division, between the scientists and engineers who produce the information and who are also the users of it and those who store and retrieve it. Here again unless what the one supplies corresponds closely with what the other needs there is a grave danger of duplication of effort. Thus the theme of the symposium is "A User-Supplier Dialogue" and I hope that NATO Scientists and Engineers who use and produce information will be stimulated to express their views and elaborate their needs.

AGARD Panels cover a very wide field and each Panel consists of specialists in that particular field. Most new developments need contributions from several fields and it is here that AGARD can make its most productive contribution by bringing together in joint meetings or symposia the specialists of the several fields. Here we have the specialists in documentation and computers talking to those working in other fields and I am sure that all will benefit.

WELCOMING REMARKS

by Dr Th. Benecke

President of the Bundesamt für Wehrtechnik und Beschaffung in Koblenz
and National Delegate to AGARD

Ladies and Gentlemen,

I have the honour and the pleasure to welcome you on behalf of the Minister of Defence and in my capacity as National Delegate to AGARD.

We would like to thank you for having accepted our invitation to hold the joint meeting of the AGARD Avionics and Technical Information Panels on "Storage and Retrieval of Information" here in Munich.

As done on previous occasions by the "Wissenschaftliche Gesellschaft für Luft- und Raumfahrt (WGLR)", the meeting has been organised this time by the new "Deutsche Gesellschaft für Luft- und Raumfahrt" (DGLR) which has been set up, combining the WGLR and the "Deutsche Gesellschaft für Raketentechnik und Raumfahrt" (DGRR). I should also like to welcome you on behalf of this association.

The central location of the Künstlerhaus, in which this meeting is taking place, offers you a good opportunity to visit, in addition to your programme, the attractions of Munich and to appreciate the beauty of the city. I would like to mention here the reception to be given by the Mayor of Munich at the Town Hall and to which you are all cordially invited.

Your meeting is devoted to a very important topic. Ever increasing scientific literature has made the handling of documentation a prominent problem, and we can only hope that modern technical equipment and procedures will help us to master this problem in the future. No doubt the planned presentations and the discussions will help to define arising difficulties and to find ways and means leading to practical solutions.

I wish you a successful meeting.

PAPER 1

COMMUNICATION AND SECRECY IN SCIENCE

by

R. Schrader

Ministry of Defence, Germany

SUMMARY

The increasing volume of material published is giving rise to the use of computers in assembling, processing and indexing of research results. Experience shows that however modern the computer it can never make critical judgement or interrogation of the data it handles. Direct contact (by correspondence, discussions, conferences) will continue to play a major role in communication.

Secrecy, whether military or industrial, is always to be deprecated from the viewpoint of scientific research and should never be applied to pure research and broad fields of basic knowledge. Secrecy can only bring short term advantages.

Finally, recommendation made by the NATO Science Committee on exchanges of scientific information, co-operation in publication and co-ordination of documentation centres are recalled.

COMMUNICATION AND SECRECY IN SCIENCE

R. Schrader

1. INTRODUCTION

Modern science places an ever-increasing demand on the communication of its results, and one of the most serious problems of today is that of providing the professional worker with speedy access to the specialized knowledge in his field of interest. In fact, no research and development programme can nowadays be initiated and carried out with some hope for success, without having available good documentation. But there is also a great danger that science as such fragments into a number of independent disciplines which bear little relationship to each other, unless we are able to cope with the "information explosion". Advancement in the science and technology of documentation had been slow until relatively recently, but with the mechanization and automation of a great many of its processes, progress is now fairly rapid. Against this background, the AGARD Panels for Avionics and Technical Information agreed to jointly organize a symposium devoted to the subject of "Storage and Retrieval of Information". When the planning for the symposium began, it was noted that several conferences on much the same subject had been held during the last years, and the question was raised as to whether there was indeed a need for another meeting of this kind. On examining the programmes of these conferences, however, it was realized that, in the main, they had focussed attention on those problems which are of primary interest to the documentalist. In fact, none of these conferences had offered much opportunity to the scientist, as the user of scientific knowledge, to discuss his needs in the fields of information processing and dissemination. For this reason, both AGARD Panels eventually reached agreement to propose a programme which should fill this gap and should serve both professional groups in a joint meeting.

As one may see from the programme, the symposium intends to demonstrate how modern storage and retrieval concepts assist the documentalist in the handling of scientific information, and scientists in the audience are expected to express their views on a number of newly proposed designs and methods. But the scientists are also invited to the meeting as active contributors. I know that in a number of countries great efforts are being made to introduce the modern techniques of data handling into the traditional fields of documentation and that promising results are expected in the not too distant future, including automatic reading and language translation, based on the application of modern computers.

Hence, the success of the symposium will depend on the co-operation and joint participation of both professional groups. To this end, invitations for the symposium have gone out to all AGARD Panels and other scientists interested in the subject, and I note with great appreciation how many of them have come, in addition to the great number of documentation experts.

As I have been invited to represent at this meeting the scientist's point of view, I intend to discharge my task by discussing communication and secrecy in science. In so doing, I recognize the important rôle documentation plays in the general process whereby science advances and our scientific knowledge grows.

2. EARLY HISTORY OF DOCUMENTATION

Documentation appears to be almost as old as man's interest in science and literature, and among the earliest collections of manuscripts were archives attached to temples and

palaces. The great philosophers and scientists of antiquity were active collectors of volumes, and Aristotle was the first man known to possess a collection worthy of the name "library".

For many centuries, the library at Alexandria represented the intellectual centre of the ancient world. It is difficult to give a precise figure, but it can be said that the number of manuscripts collected from all parts of the world and available in the library was very large; however, it should be kept in mind that the papyrus roll of antiquity usually contained less written matter than a modern book. The catalogues and classified lists issued by the library were among the earliest experiments in bibliography.

Although the Romans made excellent use of technology, they did not particularly encourage scientific advancement. As a matter of fact, the Romans were not scientifically imaginative, whereas they demonstrated outstanding capabilities in such fields as jurisprudence, political and military affairs. Science, therefore, remained Greek in nature and spirit under Roman domination, and it was not until the last century of the republic that we hear of libraries in the capital. These libraries ceased to collect Greek writings when in the year 330 the capital was removed to the Bosphorus. Eventually, the aggressions and intrusions of the Germanic tribes swept away the classical learning from Italian soil and, with the fall of Rome in 476, the ancient history of documentation came to an end.

When Christian literature began to grow, libraries became part of the ecclesiastical organizations. The abbey of Monte Cassino founded in Italy about 529 was the first monastery in which a library of religious works was established, and this custom rapidly spread to all parts of the world. Monks began to write, and the use of vellum instead of papyrus resulted in the replacement of the ancient roll by the bound book as we know it today.

The Renaissance brought about a gradual broadening of man's intellectual horizon and an ever-growing desire for the collection of manuscripts and books outside the monasteries. To satisfy this desire, large public libraries were created in France and Italy. When in 1453 Constantinople was conquered by the Turks, a full stream of Greek scholars began to flow into Western Europe, and the impact of Greek philosophy on Latin Christianity produced a powerful surge of intellectual activity and a keen interest in science. This development created a high demand for more and more books, and it is one of the miracles of history that the printing press was invented almost at the same time as Constantinople fell into the hands of the Turks. With the enormous increase of the scientific literature, new plans for libraries had to be developed, and the modern history of documentation began.

3. COMMUNICATION AND SCIENTIFIC RESEARCH

Science and technology can only grow and blossom in an environment which provides for free communication between scientists and for the rapid transmission of scientific knowledge to the engineers who always strive for innovations and continuously wish to use new discoveries for practical purposes. Even in the fields of defence research and development, progress depends on the freedom of exchanging technical information. In fact, communication has always been a necessary part of the scientific process, and there is today everywhere in the world a growing awareness of its increasing importance for the advancement of science. The more the body of scientific knowledge grows, the greater is the need for communication; and the more communication is provided, the better are scientists able to carry out research.

Until the latter half of the 17th century, communication was by way of direct correspondence whereby scientists kept each other informed about the work they were doing. At about this time, however, scientific journals began to appear providing a much better method of communication. In addition, they provided scientists with opportunities to proclaim their discoveries to the world, in an effort to gain recognition.

Today, the vast growth of scientific activity makes it difficult for the individual scientist to keep up with the ever-increasing number of publications. So much is nowadays published that he is completely incapable of studying all papers which could be of some help to him in his own field of research. This situation is often referred to as the "crisis in communication", and I share the hope that in the not too distant future a modern computer be designed whereby the steadily growing flood of new scientific information is indexed, processed and assembled, in order to get us over all the difficulties of this crisis. But whatever computers will be capable of doing in calling a scientist's attention to a piece of scientific information which might be critical to his work, they will never be able to distinguish qualitatively between a good paper and a bad paper, neither will they answer those crucial questions which scientists are in the habit of asking. For this reason and many others, open discussions among scientists working in the same field - scientists who correspond regularly with each other and meet repeatedly at conferences - will continue to play a major part in the communication process and hence in the advancement of science.

When in former times the number of scientists was small, a scientific discovery was usually the product of one single mind and emerged at one particular moment. Today, scientists are counted in hundreds and thousands, and we are likely to find that, at any one moment, a good many of them are engaged in the same piece of research and are about to publish much the same results. As a matter of fact, there are almost always several laboratories in the world which, in the pursuit and exploitation of new scientific ideas, move along the same lines of approach, and it is well remembered how much research was done in parallel during the Second World War, when major areas of science were regulated by demands of military security and scientific communication was almost non-existing.

4. IMPOSING SECRECY ON SCIENCE

In war, but also in peace, scientific information of direct military impact must be controlled by security regulations in order to prevent its premature leakage to an enemy. Hence, this information must be withheld from the traditional channels of scientific communication by some sort of classification, and nobody, of course, would question the need that in the interest of national defence certain kinds of scientific information must be protected. In fact, military secrecy is necessary and often vital to our survival. But in a real sense, it is bad for science.

Classified research is almost always in danger of suffering in quality, as this kind of work is not exposed to scientific criticism to the same extent as research which is done in an open laboratory. Secrecy prevents the free discussions which are so important for scientific progress. Secrecy furthermore results in large parts of the scientific community being kept in ignorance, at least for some time. It should be added that scientists carrying out research under condition of secrecy are less likely to enjoy the overwhelming wealth of scientific ideas than their colleagues working in a free and open environment. And as it frequently happens that a scientist fails to realize the significance of his own findings, it becomes all the more important that scientific observations be made available to others for further studies. Indeed, secrecy and great scientific thoughts cannot thrive together.

Secrecy also plays an important rôle in industrial research. It is difficult to think of a firm which devotes substantial funds to research revealing scientific information gained at great expense to its competitors. However, it has seldom happened in the past that the leakage of scientific knowledge caused a real loss to a company. On the contrary, the free release of basic information has in almost all cases been an advantage to all firms working in the same field of manufacture.

Secret scientific information often continues to remain classified, even if it has lost its military values almost entirely, and consequently does not become available to scientists as rapidly as that in the open literature. It is therefore necessary that

methods be introduced whereby scientific information will not remain classified long after its secrecy has ceased to have military significance. In my view, the proposal that some positive action should be required to maintain classification after a certain length of time merits consideration. Today, action needs to be taken in almost all cases in order to obtain declassification, since in present security procedures classification is heavily favoured over declassification.

The access to classified information is generally regulated by the criterion of "need-to-know". While this criterion can be easily applied to information of tactical military value, it can hardly be used for information of scientific content, as nobody in this world is able to say precisely in advance which piece of scientific information may or may not be of benefit to a particular research project. Because of a narrow interpretation of the need-to-know criterion, it so happened many times in the past that information which was not available to a scientist at the right moment has made the difference between the success and the failure of a research project. It is therefore very important to ensure that the criterion of need-to-know be intelligently used and not be allowed to hamper the free flow of scientific information. But it is also necessary to set up some form of special information service whereby those scientists who are entitled to see classified information are aware of its existence in order to avoid costly and wasteful duplication in defence research to as great an extent as possible.

Classification practices adopted by most countries overlook the fact that a scientific discovery made by one scientist can never be kept secret for any length of time, as one day, the same discovery will be made by another scientist. Whenever this happened in the past to a piece of scientific information which for security reasons was classified and consequently not published in the usual way, it was in almost all cases the popular belief that the information had been stolen by espionage. In reality, however, the information was rediscovered, as nature and her laws are open to every intelligent mind throughout the entire world, and no nation may claim a monopoly to be the only one with a capability of producing scientific ideas.

What do we gain by imposing secrecy on scientific research? Obviously, the main prize is time; and this is probably all we ever gain in any scientific field, as we may expect to prolong the time it takes our potential enemies or competitors to learn what we already know.

In fact, secrecy does not play a useful part in science, and security regulations should never be applied to pure research and broad fields of basic knowledge. Although the withholding of fundamental scientific information may occasionally provide short-term military advantages, in general it is detrimental to scientific progress and for this reason should always be avoided.

5. NATO'S INTEREST IN SCIENTIFIC COMMUNICATION

Within the Atlantic Alliance, the AGARD Technical Information Panel plays quite an exceptional part in the sense that it is the only body of NATO which continuously deals with documentation and its various problems. Established by AGARD in 1953, the Panel has launched a broad programme in the fields of aeronautical publications, but has also served NATO in other fields of documentation, whenever called upon.

At the request of the NATO Science Committee, the Technical Information Panel undertook in late 1958 to study ways and means whereby the exchange of scientific information within the Atlantic Alliance could be improved. As a result of these studies, the Panel recommended in March 1959 that a documentation liaison unit be established, which should not perform functions usually provided by an efficient documentation centre but should keep in touch with research and development activities and supply information to those scientists in the NATO countries who are in need of this information. Initially, this recommendation met with great enthusiasm, but was later on considered difficult to implement, as a

documentation liaison unit of the kind envisaged by the AGARD Panel would be of little use in the case of classified information which by necessity must be exchanged between nations on the basis of bilateral agreements. Consequently, the idea of a documentation liaison unit was dropped, with the understanding however that the subject of improved information exchange among the NATO countries be studied along other lines.

In 1962, the Technical Information Panel submitted to the Science Committee a report which dealt with the exchange of scientific information in the defence field. This report pointed to the need that there should be at least one defence documentation centre in each member country of NATO and that countries without such centres should, as a matter of urgency, initiate their establishment. All defence documentation centres should be equipped with the most efficient information-handling techniques and should endeavour to arrange the automatic release of all unclassified information to the corresponding centres of the other countries. As far as the exchange of classified scientific information is concerned, the report recommended that countries having mutual interests in a given field make bilateral arrangements. But pending further improvements and as an initial step, lists of unclassified titles of classified information should be given widespread distribution.

The report of the Technical Information Panel was approved by the North Atlantic Council in June of 1963, and the governments of the member countries were subsequently invited to take such action as they deemed necessary for the implementation of the recommendations. Recognizing the significance of the issue, the Council furthermore agreed that work in the field of documentation should continue, as indeed the communication of scientific knowledge within the Atlantic Alliance is a problem area of greatest importance to NATO.

In this discussion, I have referred twice to the NATO Science Committee and its interest in documentation. Now, I should like to call your attention to the Committee's report on "Increasing the Effectiveness of Western Science", issued in the autumn of 1960. Actually, the report was written by a special study group set up by the Committee, but it reflects the Committee's attitude towards the need for accelerated scientific progress in the Western world by means of enhanced international co-operation. As one may expect, the report deals with documentation, and it appears to me to be of some significance in this context to mention briefly the main recommendations, as they describe the demands for improved methods in publication and documentation in an appropriate way.

The recommendations, as they are listed in the Committee's report, call for closer co-operation in publication practices between the chief editors of the main scientific journals; the proliferation of scientific journals should be discouraged; the activities of all documentation centres should be co-ordinated, and a single international system of indexing should be introduced; authors of scientific publications should be invited to supply, together with their papers, abstracts which are to be edited according to specific rules and to be classified in accordance with a single and unified system; the abstracts should be published immediately; experienced scientists should be encouraged to periodically review broad fields of scientific research and to summarize their results in an efficient way.

These are the main recommendations made by the NATO Science Committee when dealing with the problems of documentation some years ago. They demonstrate the importance the Committee attaches to these problems. Needless to say, these recommendations are still as valid today as at the time they were written.

6. SCIENCE AND DOCUMENTATION

The critical review of wide research fields periodically done by competent scientists serves a valuable purpose in summarizing large portions of the available scientific literature. In my view, this kind of work should be given more credit than in the past. If wisely done, it will certainly assist us in our desire to cope better with the ever-increasing volume of scientific publications.

As the selection and presentation of scientific information can only be carried out intelligently by those who originate the knowledge, scientists must nowadays devote greater efforts to these activities than before. As a matter of fact, research can no longer be regarded as being completely separated from the communication of its results, and scientists must nowadays join the professional documentalists and accept responsibilities for the transmission of scientific information to the same extent to which they bear responsibility for research.

Scientists often produce a certain amount of redundancy when they publish their work, and it frequently happens that they issue for the same bit of research several reports all of which seem identical to the documentalist. There are reasons for this, one being that this is the way whereby scientists make their work known and consequently stand a better chance of gaining prestige and stature in the world of science. The essential point about publication is that scientists should always be conscious of the right to publish, and, in my view, they should take full advantage of this right, as they have indeed something to say to the world worth consideration, whether it is of pure character in the sense that it does not relate as yet to some known field of exploitation or whether it is of a more practical nature in a given area where the possibilities of material application are already well recognized. Nevertheless, the subjection to a kind of self-control practiced by the authors of scientific reports will certainly help the documentalists to overcome at least some of their difficulties.

There is another point which is critical for a fruitful co-operation between the scientist and the professional documentalist. How should our young scientists be trained to make better use of existing documentation facilities? When I was a student years ago, very little was provided in documentation training, and we first became aware of the existence of and acquainted with the many problems involved when we started to do laboratory work on our own. As by that time we did not know how to handle documentation at all, each of us began to develop his own method, so to speak, and I believe that the lack of proper guidance in this field is often the reason why so many scientists today prefer to organise their own documentation rather than to rely on the official information services. There is certainly a great need for improvement, and I sincerely hope that in the future universities will offer their students better training in documentation.

In conclusion, I should like to say that in my judgement the key to the solutions of the great many documentation problems lies in some kind of co-operative approach, involving the symbiotic activities of scientists and documentalists. Both professional groups must work together in an atmosphere of mutual respect and appreciation, and it seems to be very important to enhance their interplay. I trust, this symposium will contribute to this co-operation and will set an example for many more meetings of this type to be held in the years to come.

DISCUSSION

R.C.Wright: Do you think that in seven-and-a half years sufficient progress has been made with the NATO Science Committee's recommendation on co-ordination of documentation centres and introduction of an international system of indexing?

R.Schrader: No; although the report was submitted to the NATO council it has never received official standing in the NATO community. A recently prepared evaluation report by ten eminent scientists may be more effective.

S.C.Schuler: Does TIP activity include preparation of guidelines for authors on effective titling and abstracting of reports?

H.F.Vessey: Answering as Chairman of TIP, AGARD Specification 1 (issued 1956 and revised 1968) which is sponsored by the Panel, stresses the importance of these points.

D.Bosman: Does the education of scientists and technologists suffer as a result of security classification of reports?

N.N.Tanyolac: Post-graduate students doing research work suffer from not having all possible information available to them.

K.G.Schjetne: Is TIP or NATO Science Committee giving any consideration to what instruction in documentation university students should receive?

H.F.Vessey: TIP have not had discussions specifically in this matter. It is probably best for each country to devise methods of introducing instruction on documentation into courses, distinguishing between instruction in the use of documentation for students in general and the more detailed instruction for those who are intending to work in documentation centres. In this way much is already being done.

W.Spiess: In the absence of the use of a common indexing system in the NATO community, could the AGARD publish a memorandum listing the systems in use in various member countries and perhaps giving cross references between one system and another?

H.F.Vessey: This is a very difficult task and although efforts have been made to accomplish this over the years it has not yet proved possible. The LEX Thesaurus might be acceptable as a common indexing system.

D.Bosman: How does one guard security classified information stored in an automatic data processing system?

L.Feidelman: One answer is to isolate the data processing facility in a room accessible only to those with a security clearance. Classified documents can be indexed from their unclassified titles.

R.D.Kerr-Waller: Another solution is to have a small computer for the classified data only with access to the main computer. Use of codes to identify classified information in the general collection is useless because the codes can always be broken.

PAPER 2

THE SUPPLIER'S POINT OF VIEW - INTRODUCTORY PAPER

by

H.F. Vessey

Ministry of Technology, U.K.
Chairman Technical Information Panel, AGARD

SUMMARY

The tasks, problems and equipment available to the documentalist or information officer are outlined. The types of information agencies: libraries, documentation centres, information analysis centres and referral centres, and the services each provides, are briefly described.

The economic need to link information retrieval by computer with other activities such as preparation of abstract journals, indexes, SDI, is stressed.

Future developments in photographic methods of storing text, preparation of reports on tape, machine reading of texts, remote display, rapid printing techniques and education of documentation service users are discussed.

THE SUPPLIER'S POINT OF VIEW - INTRODUCTORY PAPER

H.F. Vessey

1. GENERAL

The field of information storage and retrieval requires definition, for it can extend from battlefield surveillance, to the booking of seats by an airline. The theme of the papers to be delivered at this symposium is mainly documentary, that is the storage and retrieval of documents, such as reports, which contain the information requested. However data, such as physical constants or properties of materials will also be included.

As Dr Schrader has said there have been a number of International symposia on this subject but the emphasis has been mainly on technique. Here it is hoped that the "customer" will take precedence and that the Scientists and Engineers will tell us what is required. I hope too that the computer engineers and documentalists will concentrate on "Service to the Customer" rather than lose us in the thickets of programming or the forests of indexing.

My paper is intended to form a background to the specialist papers that follow, to fill in some of the gaps caused by the limitation in numbers of speakers and finally to outline the task, the problems and the equipment available to the supplier, that is the Documentalist or Information Officer.

2. HISTORICAL

Dr Schrader has outlined the history of documentation but I should like to draw your attention to the fact that in about 500 B.C. there was a library of 10,000 documents at Nineveh which seem to have been systematically arranged and catalogued. "Documents" is perhaps the wrong description for these were clay tablets and must have presented some interesting storage and retrieval problems where mechanisation might well have been appropriate. The earliest reference to mechanisation the author has managed to trace is a short description in "Gullivers Travels" (circa 1700) of a machine to allow the writing of learned theses by the random selection of words and phrases.

3. N.A.T.O. DEFENCE DOCUMENTATION CENTRES

The task of the National Defence Centre in a N.A.T.O. country is to make known and supply to the country's scientists and engineers the unpublished and sometimes published literature in their field.

This is done in several ways:

- (a) By supplying to individual scientists the new literature in their fields.
- (b) By issuing accession lists, or,
- (c) By preparing Abstract Journals of new material.
- (d) By supplying relevant reports in response to an enquiry.

Further, the National Defence Documentation Centre has the responsibility of making known, in other countries, the work of its own scientists.

The broad lines of the operations are as follows:

3.1 Acquisition

The centre receives from National Establishments, Universities and from abroad a wide collection of reports and in addition may scan published material.

3.2 Recording, Abstracting and Indexing

These are processes essential to any large organisation in order to allow of retrieval. A growing number of reports, in line with A.G.A.R.D. recommendations, now contain author abstracts and even if these are not used directly they now require little editing. Indexing to allow retrieval, is a problem which will be discussed later (paragraph 7).

3.3 Translation

Most Centres will require translation of some of the foreign reports and this frequently causes difficulty as technical knowledge, as well as linguistic skill, is required.

3.4 Announcement

All Centres prepare and circulate lists of books, journals and reports received. In the case of the larger ones these usually take the form of Abstracts Journals where the new material is listed under subject headings and an abstract is given.

3.5 Circulation

Circulation procedure varies but typically a Centre distributes both at home and abroad according to agreed lists which may be standardised but which are generally unique to each report. The reports it received from outside will be circulated to a smaller circle of its own scientists either by lists or by a knowledge of particular interests.

3.6 Requests

Requests may be for a report given as a reference or listed in an Abstract Journal in which case a quick clerical search to guard against transposition of numbers, etc. is adequate. More difficult are requests for reports on a specified subject. Here a subject search is required and the success depends on several factors of which the more important are:-

1. The detail in which the enquirer can state what he wants.
2. The skill of the Information Officer in converting the request into the index terms.
3. The thoroughness of the original indexing.
4. Finally the mechanics of the retrieval process.

Finally there is the question of search for data. The documentalist will make a subject search to turn out literature in which the data may be expected to be recorded. If this is not successful he will probably attempt to put the enquirer in touch with a specialist in the field. In most organisations data searches are difficult, costly and not altogether satisfactory and there is now a tendency to concentrate this type of work in specialist centres (see paragraph 5).

3.7 Bibliographies

Bibliographies may be prepared as a result of a subject search or from known interests in a specific field. They may or may not include abstracts but should preferably be arranged in an ordered manner by subsidiary subject.

3.8 "Security" Control

Circulation, announcement and loan of reports must be controlled so that copies are only sent to those entitled to receive them. Restrictions include Security Classification (Confidential, Secret, etc.) but it should be noted that many reports marked "Unclassified" may not be widely distributed because of proprietary rights, Patent questions or policy. Such Unclassified reports which must have limited distribution should obviously be marked appropriately and there is now a growing tendency to mark the remainder "Unlimited" to indicate that there are no restrictions on circulation.

3.9 Exchange

An important activity is the exchange of reports with other similar Centres. Formal and informal exchange agreements are made, particularly with Documentation Centres in other countries. It is here that A.G.A.R.D. has made a major contribution, both in encouraging exchange and in bringing together the Heads of the National Documentation Centres in the Technical Information Panel where common difficulties may be frankly discussed.

4. OTHER CENTRES

In N.A.T.O. countries there are a number of documentation units or Agencies apart from the National Defence Documentation Centres. Each large research establishment will have its own library or Information Agency and there will be the civil technical libraries attached to Universities, Research Associations, etc. Methods of operation will vary but the basic requirements as previously described apply but with a modified emphasis. It is highly desirable that there shall be collaboration between all these areas of documentation.

5. TYPE OF AGENCY

The types of Agency may be described under four main headings although there is usually considerable overlap of activities particularly in the first two.

5.1 Libraries

Libraries deal with published information in the form of books, periodicals etc. In most cases the records used for retrieval consist of the bibliographic information (Title, Author and Publisher) and a broad statement of the main subject.

5.2 Documentation Centres

The main task is the collection, announcement and dissemination of unpublished report literature although published literature may also be processed. The records are usually kept in a more detailed form and most Centres will make subject searches, prepare bibliographies etc.

5.3 Information Analysis Centres*

These work in a specialised field, have direct access to working scientists in that field and will provide data, advice and evaluation in addition to answering detail subject enquiries.

*Sometimes called Specialised Information Centres

5.4 Referral Centres

Referral centres accept enquiries and refer them to the organisation best fitted to reply.

National Defence Documentation Centres fall within the second category but may extend into the other fields.

6. MECHANISATION

Mechanisation is increasing rapidly and many Agencies are now using computers in their operation. We shall hear of some of these in the later papers but a few words on the need to mechanise are appropriate here.

Many documentary units are running very efficiently on manual systems and to maintain balance we shall hear of the operations of one or two. There is a very real danger that a manual system that is not operating satisfactorily will be automated to improve its efficiency. The result is unlikely to be satisfactory for a poor manual system will, when put on a computer without reorganisation, be even less efficient and will cost some ten times as much.

In an organisation a case for a computer is frequently made not on the basis of retrieval but on the other operations such as the preparation of Abstract Journals, Indexes, a considerable number of Bibliographies or other such lists and finally on "House-Keeping". House-Keeping will include circulation control, "security" check, stock control and other similar functions. Retrieval will certainly be done, using the information fed in for the other operations but this alone is unlikely to be financially attractive unless the operation is very large indeed or is a "Selective Dissemination of Information" (S.D.I.) operation with a considerable number of customers (S.D.I. of course is a specialised retrieval search).

7. INDEXING

Indexing is one of the most difficult problems in documentation for unless a paper is competently indexed it is lost in the system. For some years subject specialists will be required in documentation centres for this work. Good men are scarce and expensive but as they are also required for abstracting and circulation recommendation the indexing is but a small addition to the load. As abstracting becomes less necessary and circulation is taken over by S.D.I. the pressure to develop mechanised indexing will increase.

7.1 Mechanical Indexing

Much work is being done on this subject but nothing has yet been demonstrated which is suitable for large collections. KWIC (Key word in context) Indexes where the significant words of a title are taken in turn, sorted into alphabetical order and printed out with the remainder of the title are very good for current awareness but fail for large collections. Other programmes are available which give a more attractive layout but suffer from the same defect that the resultant indexes are very much diluted by non-significant words and that the words of the title may be a poor description of the subject. The programmes certainly eliminate a number of non-significant words but the problem remains that a word may be significant in one title but useless in another. Titles now have more meaning but the author recently had a microfiche where the title of the report was "Fuzzy Sets". N.A.S.A. in STAR does rewrite titles as "Informative" titles and this is a considerable improvement and it would be interesting to hear whether any experimental work is being done in using these titles for retrieval.

A word count has been proposed. The words of the abstract or of the reports are surveyed and those occurring most frequently are sorted and printed out as indexing terms. Like KWIC

indexing this produces a much diluted index and on occasions fails because the significant words occur infrequently or are replaced by synonyms.

It is the author's opinion that the solution to this problem will come from the work of the machine translators, perhaps by a modification of the highly developed "dictionary look-up" procedures and word association. The solution may result in the allocation by computer, of conventional terms but the final aim should be to develop linguistic programmes so that the input may be in plain language and that the "dictionary" is used to scan, say the abstract and to obtain a match despite synonyms.

8. NEWER DEVELOPMENTS

A number of the newer developments deserve mention for not all will be known by the large range of customers of documentalists.

8.1 Information Analysis Centres

Information Analysis Centres have grown up to meet the needs for detail advice in specialist fields. We shall hear of the working of several of these and it is sufficient here to say that they are usually based on a University or Establishment working in a specialised field.

8.2 Referral Centres

Referral Centres have rationalised the practice of most documentary units of referring some enquiries to other agencies better able to assist. They accept enquiries, transfer them to the Unit working in the appropriate field and monitor the results.

8.3 Selective Dissemination of Information

Selective Dissemination of Information is not new and a large number of Libraries and Information Units keep a "field-of-interest" register which is used to route the new material. However, S.D.I. is the term now applied to the large operations which are now possible by computer of which we shall hear. Basically S.D.I. is a retrieval operation but the questions asked of the computer are "profiles" of each customer defining his interest. The "profiles" have to be built up by allocating index terms so that the search picks out the reports likely to be of interest to the particular scientist. Difficulties arise in establishing and, in particular, keeping profiles up to date and experience seems to indicate that the profile of a group of scientists produces more satisfactory results than individual profiles.

8.4 Citation Indexing

Citation Indexing is comparatively new and is a computer operation producing an index of all papers in which reference is made to one older publication. Thus a scientist knowing one authority in the field in which he is interested can turn up the author and title of all later papers which give the original report as references. The process is expensive in preparation and time consuming in use for not all references will be on the main subject but in some cases it will produce papers missed by other searches. It is certainly an additional bibliographic tool which in some cases will be very valuable indeed. However, despite some claims, it cannot replace other methods of search.

8.5 Microform

Photography has been used for many years in relation to storage of documents. Records on film, 35 mm, 16 mm or cut film have been made to duplicate records for preservation or for easy transmission. Some systems too have been developed for storage and retrieval. The important recent development has been the widespread use of microfiche particularly in U.S.A.

The standard American fiche is a sheet of cut film 6" x 4" containing 60 images of approximately A4 size documents but the older European systems generally used smaller fiche. Microfiche have the advantages of ease of storage, postage and reading with relatively cheap equipment. At present however, particularly in Europe, printing hard copy from them is difficult and expensive and there appears to be more resistance from scientists to using them in a reader rather than asking for hard copy. The microfiche images are at a scale of 1-18 or 21 but much greater reductions are now being demonstrated. Any documentary centre of reasonable size should have facilities for reading film and microfiche and for taking paper copies of selected pages, that is they should have reader printers. The cost of the printing of hard copy and the preparation of microfiche is still high and these tasks, for some time, will probably be confined to a few centres in each country.

9. THE FUTURE

We shall hear of several new proposals which will influence the future and the author will, therefore, merely set down those developments which he considers of major importance.

9.1 Preparation of Reports on Tape

Apart from speeding up the issue of reports which require re-drafting the final tape can be fed directly to a computer for documentary processing. At the present time an expensive process of editing and key punching is required to extract the bibliographic data and present it to the computer.

9.2 Text Reading

Although more expensive than the above, machine reading of text typed on a standard form in special type face will show advantages over key punching for those reports where tape is not available.

9.3 Time Sharing

Time sharing on computers by making use of "dead" computing time occupied by, say print out, allows a number of unconnected tasks to be done almost simultaneously. As a further extension tasks may be fed to the computer by telephone from a number of remote terminals and the coded tasks stored until computation time is available. We shall hear of one or two of the documentary applications.

9.4 Display Techniques

Display techniques are advancing rapidly and it is already possible to envisage the remote display of, say, abstracts in response to a subject enquiry.

9.5 Rapid Printing

On a more mundane level there is a very real need for economic printing at a speed compatible with computer speeds. This and display techniques will no doubt progress together for remote printing while computer typesetting is already in limited use.

9.6 Tape Exchange

It is highly desirable that documentary centres shall be able to exchange tapes and so avoid the duplication of processing that is now required. Unfortunately this is seldom possible due to incompatibility of documentary format, indexing or machine language (most probably all three). The problems of obtaining agreement are considerable but as they are not of interest to the customer they will not be enlarged on here. It should be sufficient to say that several international bodies are taking this subject very seriously indeed and much effort is being devoted to it.

Finally, it should be noted that for many years there will still be the need for local information centres without much mechanisation, although there will probably be a tendency for them to rely more and more on the large fully mechanised central stores.

10. EDUCATION

The need for education of Information Officers or Librarians in the operations and in the customers' needs is now fully recognised and the recommendations made by the several bodies considering their problems are now being implemented in most countries. Less well recognised is the need to educate the customers in the many sources of information that are now available and the best ways of using them. The author being a scientist (and engineer) converted to documentalist is acutely aware of the deficiencies on both sides and hopes that the talks on this subject will do something to remove the misunderstandings which still exist.

Management also requires education for although a few organisations fully recognise the importance of documentation they are in the minority. As an example of good practice one firm appoints an Information Officer to the team working on a new project. He attends the meetings and is responsible for searching the literature and feeding the team with all the appropriate information that is available.

The Information Officer is anxious to assist the scientist but his task is made easier and the result will be better if the scientist is aware of the documentary problems. Some specific recommendations are:-

- (a) In report writing use an informative title and include an abstract of some 100-200 words which fully describes the report.
- (b) In making requests for documents be as specific as possible. Thus where the request arises from a reference, quote originator, Report number and say author. This gives sufficient redundancy to allow for a check against, say, transposition of numbers. If the information available is incomplete, give everything there is and say "this is all available".
- (c) In asking for a subject search, be as specific as possible and give an indication of whether the search can be limited by date. If it is to be a detailed search for obscure information give some advice on the type of report which may contain it.

In case it should be felt that these recommendations are too elementary it should be said that in the author's organisation some 15% of requests are incomplete and either require a search on inadequate data or reference back. The delay, nuisance value and increase of costs is quite considerable. Again, T.I.P. was asked to prepare a bibliography on "Brittle Materials" and it was not until the work had started that the author, by approaching the British Panel Member, ascertained that interest was only in the Ceramics.

In concluding I should emphasise that this is an introductory paper and that some of the subjects I have touched on will be dealt with much more fully by later speakers. I hope however that I have given an overall picture of the activities of the several types of Information Centres, the problems which face the supplier of information and some of the new tools which are becoming available to him.

DISCUSSION

R.Bree: Could you comment on the possibility of microstorage of full texts within the computer store?

H.F.Vessey: Methods of dense storage on microform have been demonstrated with reductions of 200 : 1 and this could be one method. Another is to store information on magnetic tape and use this to generate a cathode ray display.

C.A.Bell: Are any details of the McGraw Hill Memory Bank on Defence Projects available?

W.C.Christensen: The McGraw Hill system is probably the SWEETS' System. A much better source of information on continuing U.S. research and technology is the Smithsonian Institute's data bank on research and technology.

PAPER 3

LES PROBLEMS POSES PAR LE
VOCABULAIRE DOCUMENTAIRE ET L'ORGANISATION
DES DICTIONNAIRES ET THESAURUS*

par

F. Levéry

IBM, France, Paris

avec

APPENDICE

LE CENTRE DE DOCUMENTATION DE LA
CIE IBM FRANCE, CENTRE D'ETUDES ET
RECHERCHES, LA GAUDE

par

R. J. Dubon

IBM France, La Gaudé

* Présenté à Munich par M. Dubon

SUMMARY

The compilation of documentary card indexes implies that the information contained in a text can be characterized by means of a certain number of signs. The overall signs used and available make up a documentary vocabulary serving two purposes: on the one hand, characterizing texts (analysis of subject matter); on the other hand, expressing documentary research. To fulfil both tasks, documentary vocabulary must be extensive (accuracy of representation) and structured (research strategy). Structured documentary dictionaries, or thesauri, appear to be the indispensable link between authors and people requesting information. A few practical methods for achieving such thesauri are presented.

RESUME

La creation des fichiers documentaires suppose que l'information contenue dans un texte puisse être caractérisée à l'aide d'un certain nombre de signes. L'ensemble des signes utilisés ou disponibles constitue un vocabulaire documentaire utilisé à deux fins: d'une part, la caractérisation des textes (analyse du contenu), d'autre part, l'expression des recherches documentaires. Ces deux fonctions exigent que le vocabulaire documentaire soit étendu (précision de la représentation) et structuré (stratégie de recherche). Les dictionnaires documentaires structurés ou thesaurus apparaissent comme la liaison indispensable entre les auteurs et les demandeurs. On décrit quelques méthodes pratiques permettant d'obtenir de tels thesaurus.

LES PROBLEMES POSES PAR LE VOCABULAIRE DOCUMENTAIRE ET L'ORGANISATION DES DICTIONNAIRES ET THESAURUS

F. Levéry

1. LES BUTS D'UN VOCABULAIRE DOCUMENTAIRE

Le principe de tout fichier documentaire consiste à représenter le contenu des documents, c'est à dire l'information présentée par l'auteur, à l'aide d'un certain nombre de signes ou codes. L'ensemble des signes susceptibles d'être utilisés, constituent un vocabulaire documentaire.

Dans le fichier, la totalité de l'information d'un texte se trouve réduite à la seule information contenue dans les signes ou codes qui ont été retenus au moment de l'indexage du texte. Il s'ensuit évidemment une certaine perte de l'information puisqu'un certain nombre de notions présentées dans le texte, sont exclues de la représentation documentaire.

Si l'on analyse ce phénomène d'une manière un peu plus précise, on s'aperçoit que la perte d'information au moment de l'indexage peut être due à deux causes distinctes:

- ou bien cette perte est *volue*, c'est à dire que l'analyste a considéré que certaines informations contenues dans le texte n'étaient pas suffisamment importantes pour les faire apparaître dans le fichier. Le documentaliste limite volontairement la profondeur d'indexage et ne conserve que les notions qu'il juge nécessaires. Cette limitation peut être due soit à une spécialisation poussée du fichier documentaire, (ce qui amène à exclure des notions sans rapport avec la spécialisation) soit à des contraintes extérieures (temps, volume à mémoriser, etc...).
- ou bien la perte d'information est *subie*. Ceci se produit lorsque le vocabulaire documentaire mis à la disposition de l'analyste ne permet pas de prendre en compte l'information présentée par l'auteur. Deux raisons essentielles peuvent être à l'origine de cette impossibilité:
- la notion exprimée par l'auteur n'a pas d'équivalence dans le vocabulaire documentaire. Il s'agit alors en général d'une notion nouvelle (donc malheureusement importante du point de vue documentaire):
- le vocabulaire n'est pas suffisamment précis ou étendu. La représentation de certaines notions n'est possible qu'en faisant des approximations: on utilise souvent des codes du langage documentaire correspondant à des notions plus générales.

Nous voyons donc que la perte d'information dans un système documentaire est un phénomène inévitable, inhérent à la fonction même de l'indexage. Il convient cependant de mettre au point des langages documentaires tels que cette perte d'information puisse être contrôlée. Il faut que l'analyste puisse fixer lui-même la perte tolérable en précisant la profondeur d'indexage nécessaire. Il faut éviter que cette perte soit la conséquence d'un langage documentaire insuffisant ou mal adapté.

Les réflexions qui précèdent ont amené les documentalistes à considérer qu'un bon vocabulaire documentaire devait avoir deux qualités essentielles:

- le vocabulaire doit être *precis*, de telle manière que toute notion, aussi spécifique et détaillée soit-elle, puisse trouver la représentation exacte.

- le vocabulaire doit être *extensible* de façon à prendre en compte des notions nouvelles qui apparaissent dans les disciplines techniques et scientifiques en évolution.

La nécessité de réunir ces deux caractéristiques a souvent entraîné l'abandon des systèmes traditionnels de classification hiérarchisée. Ces systèmes, en effet, ne permettent souvent qu'un indexage à l'aide de notions générales et sont trop rigides pour s'adapter à l'évolution des sciences et des techniques.

Les deux qualités de précision et d'extensibilité se trouvent en fait réunies dans le langage naturel utilisé par les auteurs. En effet, le fait qu'une notion, aussi précise ou nouvelle soit-elle, ait pu s'exprimer dans un document, montre que le vocabulaire naturel contient les termes susceptibles de l'exprimer. L'ensemble des mots utilisés par les auteurs semble donc constituer un vocabulaire documentaire ayant les qualités requises.

Il serait certainement possible d'objecter que l'information contenue dans un texte n'est pas représentée à l'aide du seul vocabulaire, mais également à l'aide de relations logiques et syntaxiques. La représentation documentaire d'un texte à l'aide d'une suite de termes du vocabulaire naturel (mots-clés ou descripteurs) non syntaxiquement reliés entre eux apparaît donc comme insuffisante. Nous ne contesterons pas cette objection. Nous y répondrons seulement en disant qu'il est nécessaire de mettre en oeuvre une certaine syntaxe lors de la création des fichiers documentaires utilisant le langage naturel. Nous ne développerons pas ce point plus avant, quel qu'en soient l'intérêt et la gravité, nous réservant de traiter uniquement des problèmes posés par l'établissement des dictionnaires documentaires.

2. CONSTITUTION D'UN VOCABULAIRE DE MOTS-CLES

S'il est exact qu'un analyste trouve parmi les termes utilisés par l'auteur d'un texte, tous les termes nécessaires à la représentation documentaire de ce texte, la constitution d'un dictionnaire documentaire apparaît simplement: ce dictionnaire sera constitué par l'ensemble des termes retenus au cours de l'indexage de tous les textes d'une collection. Il se constitue au fur et à mesure que les textes sont analysés. Le dictionnaire apparaît ainsi comme une *conséquence* et non un préalable de l'indexage. C'est un sous-produit de l'indexage.

D'un point de vue pratique, cette manière de constituer le vocabulaire documentaire est très intéressante. On est en effet sûr que tous les termes de ce vocabulaire sont utiles. D'autre part, la création du fichier documentaire peut être entreprise sans attente, ce qui était impossible avec les systèmes traditionnels puisque l'indexation des documents ne pouvait commencer qu'après la construction d'une structure classificatoire.

Il semble, d'après ce qui précède, que l'importance du vocabulaire documentaire, mesurée en nombre de termes, dépendra de deux facteurs: le nombre de documents de la collection et le profondeur d'indexage, c'est à dire le nombre moyen de termes retenus pour chaque document.

L'expérience montre que ceci n'est pas exact.

La probabilité pour qu'un terme nouveau apparaisse, diminue avec le nombre de documents précédemment analysés.

Au bout d'un certain temps, le nombre de termes nouveaux devient pratiquement négligeable. Ces termes, en général des néologismes concernent des notions nouvelles. On notera au passage, l'intéressante utilisation de cette méthode pour détecter d'une manière automatique l'apparition d'un concept nouveau dans une discipline.

Lorsque le dictionnaire ne s'accroît pratiquement plus, on dit que l'on a atteint un 'vocabulaire terminal'.

Reste à savoir comment ce vocabulaire terminal varie en fonction de la profondeur d'indexage.

Ici encore, l'expérience montre que la profondeur d'indexage n'influe pas sensiblement sur l'étendue du vocabulaire documentaire et ceci s'explique aisément:

Pour une profondeur d'indexage donnée, le fait, pour un terme, d'avoir été exclu de la liste des mots-clés, ne dépend pas de ce terme en lui-même, mais seulement de son importance plus ou moins grande dans le texte. Ce même mot pourra être considéré comme mot-clé pour un autre texte où il jouera un rôle plus important. Dans ces conditions, le vocabulaire terminal sera le même, quelle que soit la profondeur d'indexage. On l'atteindra seulement plus ou moins rapidement selon que le nombre moyen de mots clés utilisés par document sera plus ou moins élevé.

En fait, le vocabulaire terminal apparaît comme caractéristique d'une discipline. Cette remarque est particulièrement importante puisqu'elle assure la compatibilité et même l'identité des différents vocabulaires utilisés par plusieurs centres documentaires traitant de la même discipline mais avec des profondeurs d'indexage différentes (Fig.1).

3. VOCABULAIRE DE L'INFORME ET DU NON-INFORME

Toutes les méthodes de recherche documentaire dans des fichiers reposent sur le même principe. il s'agit de retrouver les documents qui ont été indexés à l'aide des termes du vocabulaire documentaire qui caractérisent la demande.

Lorsque ce vocabulaire est constitué de termes du langage naturel, la demande se présentera sous forme d'une certaine combinaison logique de mots clés.

Il est bon de noter que cette manière de faire repose en fait sur une hypothèse implicite: on suppose que le demandeur a une connaissance suffisante du sujet sur lequel porte sa question pour pouvoir établir la liste des mots-clés qui vont guider la recherche.

Or, cette hypothèse n'est pas toujours vérifiée. Bien souvent le demandeur n'a que des idées imprécises (et c'est d'ailleurs la raison pour laquelle il cherche à se documenter). La terminologie correspondant à la recherche lui est souvent inconnue et sa demande est libellée à l'aide de termes totalement différents de ceux que les auteurs utilisent pour répondre à cette demande.

Une analyse systématique du vocabulaire employé par les auteurs et de celui rencontré dans les demandes de documentation montre qu'il existe en fait deux vocabulaires distincts:

- le vocabulaire des auteurs ou 'vocabulaire de l'informé', en général précis et étendu
- le vocabulaire des demandeurs ou 'vocabulaire du non-informé', beaucoup plus limité et constitué de termes plus généraux.

Une vérification faite dans un centre de documentation sur l'électronique a donné les résultats suivants:

L'analyse des documents avait fourni un vocabulaire de 4683 mots. Les questions indexées par les demandeurs n'avaient utilisé que 1260 mots, soit environ 26% du vocabulaire disponible. Parmi ces mots, il y en avait 73 (de nature assez générale) qui représentaient à eux seuls 31% des mots utilisés pour l'ensemble des questions.

Cette différence constatée entre les vocabulaires utilisés par les auteurs et les demandeurs, rend problématique la qualité d'une sélection documentaire fondée uniquement sur la comparaison des listes de mots clés représentatifs des demandes et des documents.

On notera au passage que cette difficulté particulière n'existait pas lorsque l'on utilisait des systèmes de classification hiérarchisés. En effet, ces systèmes fournissaient facilement pour chaque notion, la liste des notions plus spécifiques, plus générales ou simplement voisines. La structure même du système de classification servait de guide au demandeur et facilitait le travail d'indexage des demandes. Elle permettait de trouver un degré de précision commun pour la représentation des documents et des interrogations, ce qui assurait la liaison entre le vocabulaire de l'informé et celui du non-informé.

Lorsque l'on utilise pour la création et l'interrogation des fichiers les termes du langage naturel la liaison entre les deux vocabulaires doit être explicitée. Il faut donc construire des dictionnaires qui permettent de trouver les termes du vocabulaire des auteurs qui se trouvent implicitement concernés par tout terme apparaissant dans une demande.

Ceci revient à dire que pour tout terme du vocabulaire documentaire, il faut établir la liste des termes qui ont avec lui une certaine relation de signification. Les dictionnaires structurés constitués par l'ensemble de ces listes portent le nom de **THESAURUS**.

L'utilisation de ces thesaurus peut être envisagée de deux manières différentes:

- une première technique consiste à se servir du thesaurus au moment de l'indexage des textes. On fera figurer dans la liste d'indexage de chaque document, certains termes qui ne figurent pas dans le document lui-même mais qui pourraient exister dans une question pour laquelle le document serait pertinent. En général cette utilisation du thesaurus à l'entrée s'accompagne d'une certaine normalisation du vocabulaire d'indexage, en particulier de la réduction des synonymes. Une telle organisation suppose bien entendu que le thesaurus préexiste et a été constitué avant l'indexage des documents.
- Une deuxième méthode consiste à consulter le thesaurus au moment de l'interrogation. On recherchera tous les termes sémantiquement reliés à ceux de la demande et qui ont pu être utilisés au cours de l'indexage des documents. Cette méthode présente l'avantage de ne pas subordonner la constitution des fichiers à l'établissement préalable du thesaurus.

L'utilité du thesaurus n'apparaît qu'au moment des interrogations une fois le fichier constitué.

Le choix entre les deux méthodes dépend en fait de facteurs économiques:

si la collection documentaire est limitée et le nombre de demandes important, il y aura intérêt à utiliser le thesaurus à l'entrée une fois pour toutes au cours de l'indexage.

Si, au contraire, le nombre de documents est très important par rapport au nombre de demandes, il sera préférable d'utiliser le thesaurus au cours de l'interrogation.

4. CONSTITUTION DES THESAURUS

Il est possible de distinguer trois types de relations sémantiques entre les termes du vocabulaire documentaire:

- *les relations d'équivalence* qui conduisent à la création de dictionnaires de synonymes
- *les relations d'inclusion* qui traduisent les relations hiérarchiques entre termes et qui correspondent aux différents degrés de généralité des notions.
- *les relations de voisinage* qui permettent d'affirmer que plusieurs termes, sans être synonymes ni dépendant hiérarchiquement les uns des autres, recouvrent un certain nombre de concepts communs.

Ces différents types de relation doivent bien entendu apparaître dans les Thesaurus, mais il existe des méthodes différentes pour les mettre en évidence.

(a) *Dictionnaires de synonymes*

L'expérience montre que les dictionnaires de synonymes linguistiques sont difficilement utilisables à des fins documentaires. Ils concernent en effet, en général, des domaines très vastes dans lesquels la synonymie est prise sous un aspect rigoureux. En documentation, au contraire, le domaine sémantique est plus spécialisé et la synonymie documentaire est plus large. Il suffit de pouvoir affirmer que tout document indexé à l'aide du mot A doit être pris en considération pour toute demande indexée à l'aide du mot B (et réciproquement), pour considérer A et B comme synonymes documentaires.

La création du dictionnaire des synonymes peut être réalisée d'une manière assez simple au cours même de la création du vocabulaire documentaire:

Pour chaque mot nouveau considéré comme mot-clé, on fait une recherche rapide, non-exhaustive de ses synonymes documentaires. On obtient ainsi une liste, forcément incomplète, des termes qui expriment la même notion. On affecte un numéro de notion à cette liste. Si, au cours de l'indexage des documents suivants ou du traitement d'une demande, il apparaît un terme nouveau, synonyme oublié de la liste précédente, une procédure analogue lui sera forcément appliquée. Au cours de cette deuxième recherche, il est parfaitement improbable qu'aucun des mots de la première liste n'apparaisse (on peut oublier un mot dans une liste mais pas une liste entière). Il y aura donc au moins un terme du vocabulaire qui sera considéré comme synonyme de deux notions distinctes et qui sera affecté de deux numéros de notion différents. Cette anomalie peut être très facilement détectée par des moyens automatiques, ce qui conduit à une correction pratiquement automatique du dictionnaire de synonymes.

(b) *Relations hiérarchiques*

La mise en évidence des relations d'inclusion entre les termes du vocabulaire documentaire peut être obtenue d'une manière relativement simple en utilisant les propriétés des systèmes traditionnels de classification hiérarchisée.

Dans un domaine sémantique donné, chaque terme du vocabulaire possède une certaine signification qui peut être rattachée à une rubrique d'un système de classification. Il suffit pour cela de considérer que la signification du terme est de nature analogue à un document que l'on chercherait à classer. Tous les termes ayant entre eux des relations hiérarchiques, se retrouveront classés sous des rubriques hiérarchiquement reliées, ce qui facilite considérablement le travail de rédaction du thesaurus.

Une autre méthode consiste à constituer le thesaurus au fur et à mesure des interrogations.

Les documentalistes qui indexent les demandes ont, en effet, l'habitude de faire éclater les mots-clés généraux et de les remplacer par une liste de mots-clés plus spécifiques reliés entre eux par des OU logiques. En effectuant ce travail, ils constituent en fait des micro-thesaurus instantanés qui traduisent les relations hiérarchiques entre les mots-clés généraux de la question et les mots plus spécifiques du vocabulaire documentaire. Si l'on prend la décision de conserver systématiquement et de mémoriser ces associations, on obtient un thesaurus de relations hiérarchiques. Ce thesaurus sera d'autant plus complet que le centre documentaire sera plus interrogé. Nous voyons donc que le thesaurus peut être constitué à partir de l'analyse des demandes, d'une manière assez parallèle à ce que nous avons vu pour l'établissement du vocabulaire documentaire, qui, lui, était obtenu à partir de l'analyse des documents.

Cette méthode a été utilisée d'une manière automatique dans un centre documentaire qui possédait une vaste collection de 'profils' destinés à la diffusion sélective. On a

recherché pour chaque terme la liste de ceux qui lui étaient fréquemment rattachés à l'aide d'un OU.

Les listes obtenues constituaient en fait un thesaurus.

(c) *Relations de voisinage*

Ce type de relation était explicité dans les structures documentaires traditionnelles par des notations du type 'Voir aussi'. Elles indiquent qu'il existe une certaine analogie de signification (et non une équivalence) entre deux termes. Par exemple RESISTANCE MECANIQUE et USURE.

Ces relations sont en général nettement mises en évidence lorsque l'on utilise la technique déjà mentionnée qui consiste à ramener les termes du vocabulaire à une structure classificatoire: deux termes voisins sont classés à l'aide de la même rubrique.

Une méthode plus élaborée consiste à comparer systématiquement les définitions des termes. Si les définitions de deux mots possèdent entre elles un certain nombre de points communs, on peut en déduire que ces deux mots sont voisins l'un de l'autre. Cette méthode a été utilisée pour déterminer des relations de voisinage dans un vocabulaire médical. Chaque mot rencontré au cours de l'indexage des textes était indexé à son tour en utilisant un vocabulaire fondamental de définisseurs (d'ailleurs assez généraux) qui précisait la signification de ce mot. Il était alors possible de rechercher automatiquement pour chaque terme, tous les termes qui avaient une définition identique (synonymes) et tous ceux qui possédaient des définisseurs communs (termes voisins). Cette méthode présentait en outre l'avantage de définir des 'degrés de voisinage'. Le voisinage était d'autant plus étroit que le nombre de définisseurs communs était plus grand.

Cette notion de 'degré de voisinage' s'est d'ailleurs révélée particulièrement importante lorsque l'on a utilisé le thesaurus obtenu pour traiter des demandes de documentation. Il était en effet, possible de proposer au demandeur d'élargir plus ou moins la question qu'il posait. Il suffisait pour cela de prendre en compte les termes plus ou moins voisins de ceux qu'il avait employés pour libeller sa question.

5. CONCLUSION

Il est certain que l'emploi de moyens automatiques et l'utilisation du vocabulaire naturel ont permis de résoudre certains problèmes documentaires et ouvrent des perspectives prometteuses. Il n'en reste pas moins vrai que ces techniques nouvelles posent des problèmes nouveaux parmi lesquels la constitution des dictionnaires et thesaurus documentaires est un des plus difficiles et des plus urgents. Il ne faut pas se cacher que la résolution de ce problème exige des moyens étendus, mais nous pensons que l'importance du problème documentaire justifie amplement les efforts qui restent à faire dans ce sens.

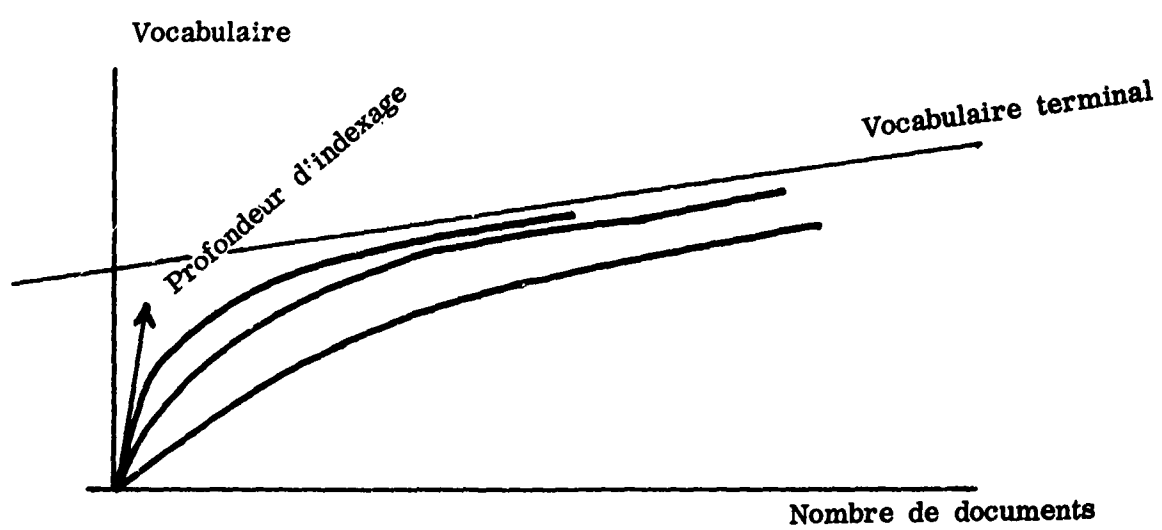


Fig. 1

APPENDIX

Le Centre de Documentation de la Cie IBM France,
Centre D'études et Recherches, la Gaude

R. J. Dubon

1. EVOLUTION DES TECHNIQUES

La mission d'un Centre de Documentation traditionnel peut se résumer à quatre tâches principales:

1. Accueillir les documents à leur arrivée au Centre: analyse, classification, emmagasinage.
2. Tenir les clients du Centre constamment informés des nouvelles parutions les intéressant directement (Diffusion Sélective de l'Information).
3. Effectuer, à la demande des clients, des recherches bibliographiques sur un sujet déterminé (Recherches Rétrospectives).
4. Etre en mesure de fournir l' 'adresse' d'un document déterminé et le document lui-même.

Dans une PREMIERE GENERATION, utilisant des méthodes manuelles, le procédé essentiellement à la CLASSIFICATION HIERARCHIQUE: ces techniques peuvent être justifiées pour des banques d'informations limitées en volume, et spécialisées en nature. A partir d'un certain volume à l'entrée, et pour des informations couvrant un large éventail d'activités, les systèmes manuels et leurs divers systèmes de classification se révèlent inefficaces, lents et coûteux.

Un premier essai d'automatisation fut tenté, à l'occasion de la SECONDE GENERATION, avec l'introduction des MOTS-CLES. Dans cette étape vers l'automatisation complète, un document technique était caractérisé par une série de mots-clés (10 en moyenne) et des références telles que: titre, date, noms d'auteurs. Cet ensemble de données était mis en mémoire de l'ordinateur puis recherché par programme. Il n'y a pas lieu de s'attarder sur cette méthode, car les résultats furent décevants: on a constaté que seulement 10% des références trouvées répondaient aux questions posées. La 'sortie' ordinateur, tout comme l' 'entrée', consistait en une liste de mots-clés pas assez informative pour décrire suffisamment le document, et par conséquent, source de confusion avec le véritable sujet d'intérêt. D'autre part, tout système utilisant classification hiérarchique ou mots-clés est dangereux, en ce sens qu'il fait appel à une analyse humaine, source d'erreurs; enfin, les technologies avancées font appel, souvent, à des notions nouvelles que seul le recul du temps permettra d'apprécier.

Ceci est grave, car un document mal analysé et indexé sera perdu à jamais lors d'une interrogation future, sauf si cette dernière est elle-même mal composée. Enfin, la nécessité d'une intervention manuelle limite le volume acceptable à l'entrée et l'on retrouve les inconvénients de la Première Génération.

Cette méthode de recherche sur mots-clés est aujourd'hui totalement abandonnée.

La TROISIEME GENERATION utilise des techniques dites de 'texte normal' ou 'langage clair', objet de notre discussion sur les vocabulaires de documentation.

2. TRAITEMENT DU DOCUMENT COMPLET A
L'ENTREE DU SYSTEME

Seul le résumé du document fait l'objet d'un traitement ordinateur. Le document complet est détruit après avoir été réduit et photographié par les soins d'IBM sur un support microfilm approprié (microfiche). Cette opération est effectuée aux USA et ne concerne pas les articles de revues et périodiques, pour des raisons de droits de reproduction.

Un jeu complet de ces microfiches est ensuite envoyé au Centre européen, ainsi qu'aux différents Services de Documentation de la Compagnie IBM dans le monde, clients du système.

3. TRAITEMENT DES RESUMES

Dans la suite des opérations, c'est donc le RESUME du document qui va faire l'objet d'un traitement automatisé et être mis en mémoire de l'ordinateur. C'est donc véritablement la "pensée de l'auteur", exprimée en langage clair, qui sera "lue" par la machine; cette dernière sera alors en mesure de "répondre" aux questions de ses clients, posées grâce à une logique d'interrogation décrite ultérieurement.

L'importance du résumé est donc très grande, car c'est sa qualité que dépendront les résultats d'une recherche.

Actuellement, la plupart des publications scientifiques et techniques sérieuses (rapports, thèses, articles de revues, etc.) ont un résumé. C'est, de plus, une règle à l'intérieur de la Compagnie IBM. En tout état de cause, notre système ne tient pas compte des documents qui n'ont pas de résumé.

L'auteur d'un document est censé être la personne la plus qualifiée pour en faire le résumé, et de ce fait le rendre le plus "informatif" possible. Le résumé comporte:

- Titre, date
- Auteur(s), origine du document
- Numéro propre du document, identifications diverses
- Résumé proprement dit, de 10 à 30 lignes (100 à 300 mots), suivi de l'indication du nombre de pages
- Un certain nombre de descripteurs (2 à 5), dont un numéro de catégorie, destinés à la préparation de catalogues par sujet, par catégorie (Liste des 23 catégories en Annexe 2)
- Un numéro d'accès séquentiel chronologique, purement artificiel, servant à identifier le résumé en ordinateur, ainsi que le support microfilm où se trouve le document complet.

Les résumés sont alors mis en mémoire de l'ordinateur par l'intermédiaire des supports "carte perforée" et "bande magnétique" au cours des opérations suivantes:

- Perforation des résumés (en moyenne 20 cartes par résumé), suivant des règles imposées par les divers programmes d'exploitation du système de Recherche Automatique de Documentation
- Mise sur bande et vérification simultanée des fonctions suivantes,
 - Validité des caractères
 - Numéro de séquence en ordre numérique croissant
 - Composition du texte clair: présence de l'ensemble des composantes du résumé
 - Orthographe: l'ensemble des mots nouveaux est comparé avec une "bande magnétique dictionnaire", contenant environ 1 100 mots correctement épelés.

Lorsqu'une anomalie est détectée - mot mal orthographié, erreur de perforation, de séquence, mot nouveau, etc - elle apparaît sur une imprimante afin de permettre aux spécialistes de la corriger; un mot nouveau est par exemple ajouté au dictionnaire sur bande, une erreur est corrigée, etc.

Lorsque l'ensemble de ces fonctions a été vérifié, des bulletins (résumés complets) et catalogues par sujet, catégorie, auteur, origine et numéro d'accès sont produits par l'ordinateur, et envoyés aux mêmes destinataires que les microfiches contenant les documents complets (Fig. 1).

Enfin, ultime étape, une série de bandes magnétiques correspondant aux résumés des nouveaux documents est préparée par type de document. Une bande magnétique à haute densité (800 BPI) contient environ 6 000 résumés de documents.

Le Centre de La Gaude reçoit alors des Etats-Unis une copie de ces bandes qui viendront s'ajouter aux bandes 'françaises'. Ces nouveaux documents sont alors fusionnés, des deux côtés de l'Atlantique, avec les fichiers 'historiques' et constituent la 'mémoire documentaire' de l'ordinateur.

N.B. - Les mises à jour sur bande sont transmises des USA à La Gaude par les unités IBM 7702 de transmission de bande magnétique, reliées entre elles par simple ligne téléphonique.

4. CARACTERISTIQUES DU SYSTEME

4.1 Entrée Simplifiée

- La pensée de l'auteur, en langue originale et en langage clair, est directement mise sur support magnétique avec un minimum d'intervention humaine.
- La recherche s'effectue sur le texte normal de l'ensemble des éléments du résumé.
- Aucune codification, aucune classification, aucun mot-clé ne sont désormais nécessaires: les résumés se trouvent sur bande magnétique, en ordre chronologique, quelle que soit leur nature.

4.2 Logique de Recherche

Souple et efficace, elle permet, également en langage clair, de consulter les fichiers documentaires sur bandes.

- Cette logique est fondée sur la satisfaction de CONCEPTS, dont le nombre et la nature décrivent le problème posé.
- Le langage-question utilise un certain nombre d'Opérateurs Logiques, permettant de définir une question, aussi complexe soit-elle.
- Les questions peuvent être posées, en clair, dans la langue du document en mémoire. Actuellement, français et anglais sont utilisés conjointement dans ce but. Cette possibilité peut être étendue à toute langue.

4.3 Rapidité de Traitement

L'ordinateur permet de poser une moyenne de 100 questions simultanément, de lire 120 000 mots de texte par minute et d'imprimer les résultats sur des imprimantes rapides.

4.4 Format des Réponses

Il est identique à celui de l'entrée, c'est-à-dire qu'il consiste en un résumé, en langage clair, avec indication de l'adresse du document complet (support microfiche).

5. LOGIQUE DE RECHERCHE (FIG. 2)

La sélection d'un résumé de document est fondée sur la satisfaction d'un "critère de sélection". Ce critère exige la présence d'un ou de plusieurs concepts simultanés à l'intérieur du résumé.

Appelons CONCEPT ELEMENTAIRE un MOT isolé du langage naturel (en n'importe quelle langue). Un CONCEPT sera la réunion de concepts élémentaires, reliés entre eux par un certain nombre d'OPERATEURS LOGIQUES.

6. DIFFERENTS CONCEPTS POSSIBLES

- 6.1 (a) Le concept le plus simple est le Concept Élémentaire. Il est possible, par exemple, d'exiger la seule présence d'un mot dans un résumé pour assurer sa sélection.

Exemple: LASER

N.B. - Il faut être prudent, car un terme trop général (transistor, ordinateur), pour un critère de sélection égal à UN (1), peut engendrer une "sortie" trop importante.

- (b) La technique du MASQUE permet de rechercher sur la racine d'un Concept Élémentaire (mot), afin de couvrir les différentes formes grammaticales possibles, en ouvrant un éventail de caractères après la racine. Deux cas peuvent se présenter:

Masque sélectif: indiqué par autant de signes § que de caractères à masquer, il permet d'ouvrir un éventail limité à 1, 2, 3, 4 ou 5 caractères après la racine du mot.

Exemple: TRANSISTOR§ couvrira le mot au singulier comme au pluriel, HOPITA§§ couvrira les mots Hôpital ou Hôpitaux.

Ce masque sélectif est utilisé lorsque l'on veut éviter la répétition d'un mot sous ses formes différentes, ces formes et le nombre maximal de caractères possible après la racine étant connus.

Masque illimité: indiqué par le signe §., il permet d'ouvrir un éventail illimité de caractères après la racine du mot.

Exemple: DOCUMENT§. couvrira des mots tels que Document, Documents, Documentation, Documentalistes, etc.

Il faut être prudent dans l'emploi du masque illimité, car la sélection peut se faire sur des mots dont la racine est commune, mais dont la terminaison peut être telle que le mot n'a plus aucun sens avec le mot de la question.

Par exemple, si l'on recherche des mots tels que ORAL ou ORAUX, ou ORALEMENT, il faut éviter d'utiliser ORA§., car des mots comme ORAGE, ORAISON, pourront être considérés, aussi, comme bonne réponse....

- 6.2 Des concepts plus complexes se forment à partir de la réunion de Concepts Élémentaires reliés par des Opérateurs Logiques.

Opérateurs Logiques:

- (a) *OU* Cet opérateur permet d'exprimer différentes possibilités, c'est-à-dire différents Concepts Élémentaires, sur le même niveau logique.
- w OU x OU y OU z
- Exemple de CONCEPT possible:* hôpita§§ OU clinique§
- (b) *ET* Cet opérateur exige la présence simultanée des plusieurs Concepts Élémentaires ou de plusieurs CONCEPTS.
- a ET b ET c
- Exemple de CONCEPT possible:* ordinateurs§ ET médecine ou encore: automatisation ET (hôpital§§ OU clinique§)
- (c) *AVEC* Cet opérateur exige la présence de deux ou plusieurs Concepts Élémentaires à l'intérieur d'une phrase de texte: une phrase est constituée par l'ensemble des caractères compris entre deux points.
- a AVEC b
- Exemple de CONCEPT:* bandes§ AVEC magnétique§.

- (d) *ADJ* Les Concepts Elémentaires affectés par cet opérateur doivent se trouver en position adjacente, et dans l'ordre.

a ADJ b ADJ c

(b doit suivre a et précéder c)

Exemple de CONCEPT: documentation ADJ automatique ou encore: mémoires ADJ à ADJ tambour ADJ magnétique

- (e) *SAUF* Cet opérateur peut s'appliquer à tout concept précédemment défini (élémentaire ou non). Il permet d'exclure de la sélection finale tout résumé dans lequel le concept non désiré apparaît, même si, par ailleurs, le résumé rencontre d'autres critères de sélection.

Exemples de CONCEPTS négatifs: SAUF (fortran OU cobol); SAUF (centres ADJ téléphoniques); SAUF (simulation ET aérospatial)

- (f) *OUI* Cet opérateur peut s'appliquer également à tout concept. Il ne s'emploie que dans deux cas:

(i) La question posée comporte déjà un opérateur SAUF. Dans ce cas, l'opérateur OUI domine et assure impérativement la sélection, quel que soit l'environnement sémantique.

(ii) Le critère de sélection exige deux concepts ou plus. Dans ce cas, une condition suffisante de sélection sera remplie si le seul concept affecté de l'opérateur OUI est présent dans un résumé. La présence des deux (ou de plusieurs) concepts réclamés par ailleurs n'est plus nécessaire.

- (g) *INDICATEURS de ZONE de RECHERCHE* Lors de la peroration des résumés, un code d'identification de chaque zone du résumé (zone-titre, zone-auteur, zone-origine, zone-résumé, etc.) fait partie de l'ensemble des caractères constitutifs de cette zone. Le but est double,

Assurer une mise en page lors de l'impression,

Permettre une recherche localisée à l'intérieur d'une zone particulière du résumé.

(i) *Indicateur impératif de zone* - S'applique à tout concept. Cet indicateur permet d'exiger la présence du concept à l'intérieur d'une zone préférentielle, indiquée à l'avance lors de la question.

Par exemple,

la question "CONTROLE ZONE AUTEUR Einstein" permettra de ne sélectionner que les documents dont Einstein fut l'auteur, et non ceux dans lesquels la théorie de la relativité d'Einstein est mentionnée.

(ii) *Indicateur d'exclusion de zone* - S'applique à tout concept. Cet indicateur permet inversement de poser une question relative à tous les termes d'un résumé, sauf à ceux d'une zone non désirée, choisie à l'avance.

Par exemple,

la question "SAUF CONTROLE ZONE ORIGINE ondes ADJ électriques" permet de retrouver les documents traitant des ondes électriques, sauf s'il s'agit des articles de la revue "Onde Electrique" (à moins que cet article ne traite précisément d'ondes électriques.)

7. CONSTRUCTION D'UN PROFIL-QUESTION

Cet ensemble de moyens logiques permet, en définissant le nombre et la nature des concepts requis, de convertir les intérêts du demandeur en langage-question clair, informatif et précis.

La construction d'une bonne question dépend essentiellement

- d'un choix correct de la terminologie caractéristique du problème posé. Pour cela, il faut une bonne connaissance du sujet, jointe à une connaissance appropriée de la nature de la banque d'informations et de son volume approximatif dans les différents domaines scientifiques et techniques qui la composent. Thesaurus et Dictionnaire de Synonymes sont également des outils efficaces pour 'couvrir' le sujet.
- d'une bonne utilisation de la logique de recherche, en fonction du sujet et du résultat recherché. Il est important de savoir si le demandeur veut obtenir peu ou beaucoup de références, afin de rétrécir ou d'élargir la question en conséquence. Le choix des opérateurs et de leurs liens réciproques est primordial.

A titre d'exemple, la comparaison entre les possibilités offertes par l'utilisation des opérateurs ET, AVEC et ADJ est intéressante.

Supposons qu'un chercheur soit intéressé par les développements dans le domaine de la technologie des bandes magnétiques. Sa question peut se définir par 2 concepts,

- (a) Le concept "technologie" et ses synonymes ne présente pas de difficultés: technolog~~ies~~ OU techniqu~~es~~ OU etc.
- (b) Le concept "bandes magnétiques" peut être nuancé par le choix que l'on fera de l'un des opérateurs ET, AVEC, ADJ.
 - (i) L'opérateurs ET (bande~~s~~ ET magnétique~~s~~) sera choisi si l'on veut une recherche aussi large que possible, puisque nous exigerons la présence des Concepts Elémentaires "bande~~s~~", "magnétique~~s~~", n'importe où dans le résumé. Toutefois, dans un tel résumé, les mots "bande~~s~~" et "magnétique~~s~~" peuvent se trouver physiquement éloignés et dans des contextes sémantiques différents, d'où un risque de "bruit", c'est-à-dire document sélectionné mais hors du sujet.
 - (ii) Si le choix de l'opérateur ADJ est fait (bande~~s~~ ADJ magnétique~~s~~), l'ordinateur ne sélectionnera un résumé que s'il comporte le mot "bande~~s~~" suivi du mot "magnétique~~s~~". Dans ce cas, nous réduisons le risque de "bruits", mais nous introduisons celui d'avoir des "silences", c'est-à-dire des résumés en mémoire, mais non sélectionnés à cause d'une logique insuffisamment adaptée au langage humain.

Supposons qu'un résumé contienne la phrase suivante,

"...Ces résultats proviennent de l'utilisation de nouveaux matériaux magnétiques récemment adoptés pour les dérouleurs de bandes IBM..."

Ce résumé ne sera pas "retrouvé" et notifié, bien qu'il réponde parfaitement à la question.

- (iii) Pour ce problème précis, le choix de l'opérateur AVEC (bande~~s~~ AVEC magnétique~~s~~) est souhaitable, car cette logique se situe à mi-chemin entre le ET, trop général, et le ADJ, trop restrictif. Dans ce cas, le résumé mentionné ci-dessus sera sélectionné, puisque les Concepts Elémentaires "bande~~s~~" et "magnétique~~s~~" se trouvent dans une phrase, entre deux points de texte.

La Charte "Logique Recherche 360", illustre les possibilités logiques décrites plus haut. Elle appelle quelques commentaires:

- les différents concepts, identifiés par le symbole CONXX, où XX représente un simple numéro de série et non de priorité, se trouvent ici sur le même niveau logique, car le critère de sélection a été fixé à 1. Chacun des 7 concepts est donc indépendant vis-à-vis des autres.

- le CON03 s'exprime par l'intermédiaire de sous-concepts, A3 et A4. Chacun des sous-concepts A3 ou A4 est insuffisant par lui-même pour assurer une sélection, puisqu'il est nécessaire de les réunir par l'opérateur ET.
- . A4 exprime des possibilités (logique OU)
- . A3 permet de réunir sur le même niveau logique, affectés de l'opérateur OU, des mots isolés (A1) et des mots adjacents (A2).

Ce concept montre toute la souplesse de cette logique, dont la complexité, tout en suivant des règles d'application strictes qui ne seront pas détaillées ici, peut être adaptée sans limites à la caractérisation du problème posé. (Figs. 3 et 4).

La discussion de ce sujet continue page 39.

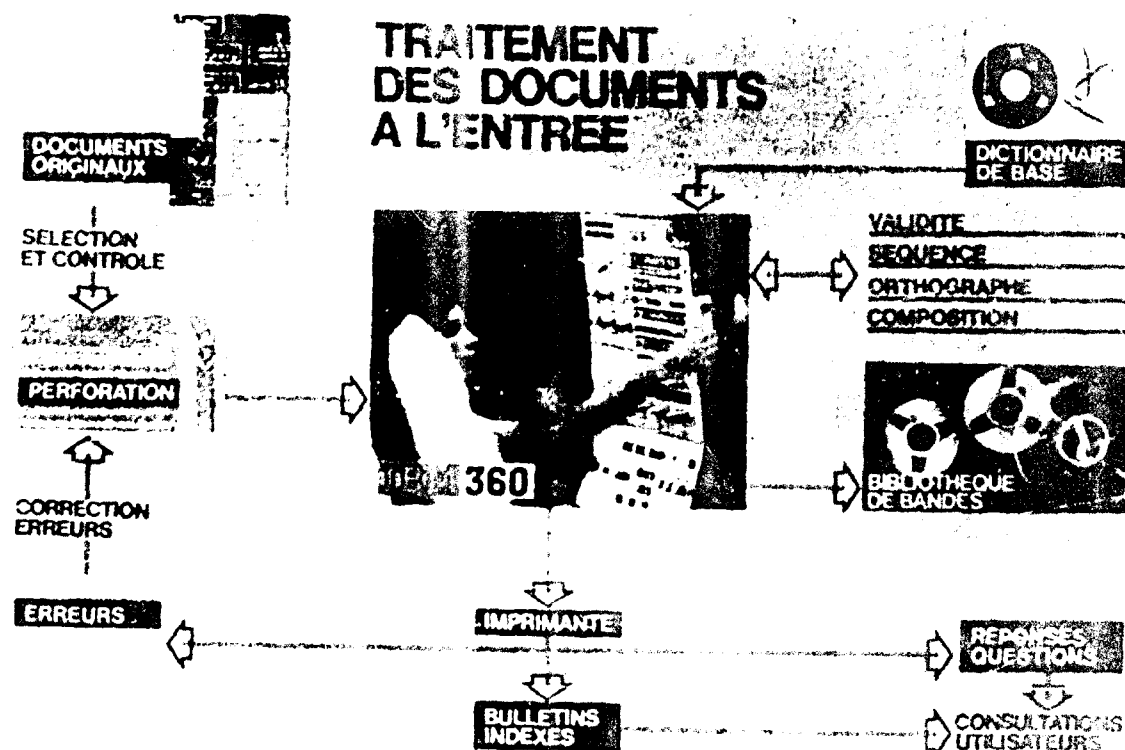


Fig.1 Traitement des documents a l'entrée

MOTS ISOLES	exige un seul mot	laser
MASQUE	recherche sur racines	transistor .
OU	différentes possibilités	hopita ... clinique .
MOTS ADJACENTS	position adjacente	circuits imprimés
PHRASES	suite de mots	mémoire a tambour magnétique
ET	X et Y et Z	ordinateur et médecine
AVEC	zone limitée	bandes. magnétiques
SAUF	exclusion	TV couleur sauf SECAM
OUI	obligatoire	Henri de France
CONTROLE	zone imposée	Einstein/ + auteur
SAUF CONTROLE	zone exclue	Onde électrique/- revue

Fig.2 Logique de recherche

CON01 INFORMATION OU DOCUMENT\$ * OU
LITTÉRATURE AVEC RECHERCHES
OU STOCK\$ *

CON02 DISSEMINATION ADJ SELECTIVE

A1 BIBLIOTHEQUES

A2 CENTRES\$ ADJ D ADJ INFORMATION\$
ADJ TECHNIQUES

A3 A1 OU A2

A4 AUTOMAT\$ * OU MÉCANISS\$ * OU CALCULAT\$ *

CON03 A3 ET A4

CON04 ABS LUHN ADJ HP

CON05 SAUF ORLEANS

CON06 ELECTRONIQUE SAUF CONTROLE.100
ADJ INDUSTRIELLE

CON07 LOUIS CONTROLE.200 ARMAND

Fig.3 Logique recherche 360

A32700 1 RJ DUBON FRAN SDD 010 CER 07\00 01

A32700 A1	SELECT\$\$\$	02
A32700 A2	DISSEMINAT\$\$\$	03
A32700 A3	DOCUMENT\$* or DATA or INFORMATION	04
A32700 Con01	A1 and A2 and A3	05
A32700 A4	RETRIEV\$* or SEARCH\$*	06
A32700 Con02	A3 with A4	07
A32700 Con03	CURRENT adj INFORMATION adj SELECTION	08
A32700 A5	LIBRAR\$\$\$	09
A32700 A6	MECHANIZ\$* or AUTOMAT\$\$\$ or ELECTRONIC\$ or TREND\$ or FUTURE or	10
A32700	COMPUT\$\$\$	11
A32700 A7	DATA adj PROCESSING	12
A32700 A8	A7 or A6	13
A32700 Con04	A8 and A5	14
A32700 Con05	NATIONAL adj INFORMATION	15
A32700 A9	DIRECT adj ACCESS	16
A32700 A10	REMOTE\$\$\$ with COMPUT\$\$\$ or CONSOLE\$	17
A32700 A11	DISPLAY\$\$\$ or VIDEO or CRT	18
A32700 A12	CATHODE adj RAY adj TUBE\$	19
A32700 A13	A10 or A11 or A12	20
A32700 Con06	A9 and A13	21
A32700 A14	A11 or A12	22
A32700 Con07	A14 and A10	23
A32700 Con08	ITIRC	24
A32700 Con09	DUBCK with RJ	25
A32700 Con10	MAGNINO with JJ	26
A32700 Con11	MERRITT with CA	27
A32700 Con12	GARLAND with J	28
A32700 Con13	JACKSON with EB	29
A32700 End		30

Fig.4 Exemple de profil

DISCUSSION

A.H.Holloway: You have described your system as using natural language, but if the natural language of the enquiry does not correspond with the natural language of the document there must be intervention by a human agent. The dialogue between the user and the system would seem to be essential to the performance of any system.

R.J.Dubon: It is true that up to the present time we have not been able to eliminate the link of the information specialist between the enquirer and the system.

E.Keonjian: What are the maintenance problems of your systems, including the error detection in the system, and the qualifications of the personnel required to operate it?

R.J.Dubon: Key punching is checked so that we know our tapes are correct. Logic errors can be detected when preparing the search program. Staff, excluding those preparing the input, consists of two engineers, a clerk and a secretary.

R.C.Wright: How many persons are engaged on input processing to the system?

R.J.Dubon: Fifteen persons are engaged on key punching, proof reading, merging and dispatching the input. No abstracting is done; the abstract accompanying the article is used.

R.D.Kerr-Waller: We have found that the use of Boolean logic in a search produces false drops. A change to a system of weighted keywords resulted in a considerable reduction in the number of false drops.

R.J.Dubon: We also have noted errors when using Boolean logic but these have not been serious. Use of a "with" logic has given good results.

R.Bree: 1. What is the total input to your system?

2. What influence does the use of several languages have on the economy of the system?

R.J.Dubon: 1. Input is 3,000-4,000 items per month.

2. Foreign languages would not affect the economy of the system as the system operates in the English language and questions are translated into English.

E.Lapeysen: Is there any screening of the output from a question put to the system?

R.J.Dubon: No; the output is sent direct to the enquirer.

PAPER 4

FOUR "NEW" SCIENCES: AN APPROACH TO COMPLEXITY*

by

E.B.Montgomery

Dean, School of Library Science,
Syracuse University, New York, USA

* Dean Montgomery was not able to deliver this paper at the Symposium but it is included to complete the Proceedings. In Dean Montgomery's absence, Dr. E.L. Elchhorn of Jet Propulsion Laboratory, USA spoke on Integrated Data Management in the Deep Space Net. As this was a shortened version of a paper intended for full presentation elsewhere, it is not thought appropriate to include it in these Proceedings.

SUMMARY

Present - day discoveries are causing an exponential growth in information because a discovery in one field may very well lead to new discoveries and the generation of new information in other, related fields and even in other disciplines. The problem is one of coping with complexity.

Four "new" sciences, not new in themselves, but looked at from new points of view are suggested as solutions, namely Information science, Communication science, System science and Application science.

Finally, it is suggested that the Science of science itself may provide answers to some of the problems.

FOUR "NEW" SCIENCES: AN APPROACH TO COMPLEXITY

E. B. Montgomery

The problem to be faced in scientific and technical information involves far more than the solution of information storage and retrieval problems of our technical culture. We are currently faced with the complexity that results from the exponential increase in knowledge. We have not learned to cope with it. And yet, these increases force us to expand our search for more knowledge.

Research and development in area after area are becoming mass produced and even automated. This trend will also increase. These increases are exponential by nature. Much, if not most, of the new information created or discovered has implications for and interrelationships with other knowledge already discovered as well as that which will be created in the future. New information frequently affects information in other areas and even in other disciplines. For example, the unravelling of the genetic coding in the RNA molecule has major implications for many disciplines. Radio carbon dating has changed many sequences of history. In turn, this changed information gives rise to more changes in other disciplines.

It is not a small, simple exponential but a complex phenomenon of chain reactions. Some of them die out quickly; however, others continue branching into many disciplines having large reproduction factors with relatively short reproduction cycles, ranging from a few months to a few years.

New approaches are needed if this complexity is to result in progress rather than confusion and chaos -- if in other words, we are to anticipate and protect ourselves from the by-products of increasing knowledge.

This paper suggests an *approach* to the solution of part of the problem. The approach calls for the organization and synthesis of "new sciences" that will allow better comprehension of the interrelationships of knowledge which can lead to such problems as pollution and inundation of information and overpopulation. Interrelationships exist in many dimensions and play many roles. The more points of view we have, the greater chance we have to understand those interrelationships of which we are already aware and to discover new and unsuspected relationships.

The four areas which should be developed into new disciplines of science are: information, communications, systems and applications.

No one of these is entirely new. They are being pursued in varying degrees at present and in various combinations. The establishment of societies for cybernetics, general systems, etc., attest to this fact. These efforts, however, are not nearly strong enough to provide what is needed in the face of the present growth of complexity.

Therefore, it is recommended that these sciences, accompanied by the obvious parallel engineering fields, be explored, structured and pursued with increased breadth and support.

At present, interrelationships sought through research and study are somewhat unilateral and are usually confined to the discipline giving rise to them. Even so, there is more

realization of their potential outside their parent discipline than ever before. But this is not nearly enough. More awareness of implications and predictability is needed.

If we postulate the four logical scientific domains of knowledge that provide a continuum from information to its ultimate use, and further require that these domains cross all disciplines at right angles to them, we should add the now missing dimension to our ability to understand the potential interrelationships of knowledge.

The first science is information science. In order to define it, we should consider the broadest possible definition of information. Whether or not the definition agrees with that in the dictionary is of little importance. Information exists everywhere in the universe around us. We *could* define information as the position of *all* the atoms and molecules in the universe and of all sets and combinations of those atoms and molecules at any time.

Let us next consider how information operates -- both as a result of man's knowledge and reflection and as the processes constantly occurring in nature. Of course, we are limited by far too little knowledge of both the thought and activities of man and of the many facets of nature. We know that we must continue our investigations which will lead to an increased knowledge about information.

Information exists within all of the organisms of nature. In fact, every cell of every organism has molecular information that defines and determines the complete description, make-up and functional specification of that organism. When organisms reproduce, information about the one or two organisms giving rise to the reproduced organism, is transmitted in the reproduction process and a new piece of molecular information is created in the process.

On the highest level of complexity, man is acted on not only by his environment and his reaction to it, but also by his own feelings and thoughts.

As a result of these complex interactions, organisms have information which is transmitted frequently, if not constantly, about the state of the organism and of the sub-organisms or systems within the organism. For example the nervous system is in constant surveillance of the sensing by the human body of the change of state of the organs, of the change of state of emotions and of the change of state of thinking.

Information is also produced by the activities of organisms. Humans observe, design experiments, create, think, feed and indulge in various activities which increase the amount of information. In a scholarly discipline, such as a science, this increase in information is sought after in an organized and logical fashion.

However, information itself is not sufficiently investigated as a scientific entity. The definition of information, its roles, and its contribution to the disciplines are more frequently by-products of disciplines with other goals than an understanding of information *per se*. Information, its dynamics, the tools that elicit it and the tools with which we use it deserve a special science that can lead us, through better understanding, to a far better use.

Information in itself, however, is of little use. It must be *communicated*. We must search scientifically for answers to such questions as: What is communication? How does it take place? Why does it occur? What are its purposes? and What forms of communication are there? There is a need for a comprehensive science dedicated to the understanding of communication within and among and between organisms.

Information communicated through a system usually has a purpose or application. Thus, a science of applications following a science of systems is needed to provide full coverage of the information continuum.

Complete understanding, then, requires a knowledge of what information is, how it is communicated, the systems used in transmittal, and, ultimately, its application. This complex is, of course, referred to by some disciplines as "cybernetic systems."

External information and communication actions of organisms include communication on a personal basis, on societal, cultural, business, international and many other bases. Among these, too, are the needs for a better understanding of what communication is, what a system is and *why* it operates or is applied.

The four sciences should not be limited to the interface with participation in the physical, biological and social sciences. These sciences of information should transcend and tie together all activities of man, including all of the disciplines, from art and architecture to zoology.

We do need to know what communication is and what a system is from a sufficiently broad point of view so that we understand their general properties and dynamics and can anticipate their by-products.

There are many research and development activities in communications and systems as well as in information science. Most of these have been organized to solve particular problems rather than to seek a broad understanding of the dynamics and theory of information, communications and systems sciences (this is at present more the object of the societies that have been formed). This work has been of tremendous value and it is not intended to detract from it.

The fourth science dealing with applications should be more thoroughly explained. For example, how does man apply the knowledge of the information that he has? Most of the endeavors in our society are those that pertain to the use of knowledge; yet one of the least perfect areas of scientific endeavor might be said to be that of applications.

A science of applications would embrace studies leading to an understanding of how man *really* uses information. This could lead to vastly improved functions of all sorts. For example, education is the application -- through a system of communication -- of the knowledge and the information man has. The end product of education is the application of knowledge to all phases of learning -- further scholarship, a professional career or simply a fuller, better life. Yet, the entire process is not very well understood. What learning is needs to be much better understood from the standpoint of the application of information. The best ways of communicating information are not known. Do we really know what learning is? The systems called schools, although they are improving, are far from achieving the best educational objectives because of our lack of knowledge of application.

Applications knowledge impinges on the arts, and while the present state of the arts may seem productive and satisfactory, one facet of art is in the ability to apply. However, what is applied and how it is applied are not very well understood.

The list of examples of the need to study applications should also include questions of how man applies his knowledge of political, economic, and social sciences to the relationships among nations. Here again, precision of application knowledge, in fact what knowledge to apply, based on what information, is an enormous problem which man has failed to solve over the centuries. Now that some sciences have succeeded in applying their knowledge to the creation of devices with the capability to eliminate man from the face of the earth, our inability to discover, to communicate, to construct fail-safe systems of society lead to the realization that our application of man's knowledge needs even greater depth and surer progress.

A fifth science, which is not a new science at all, might be called the science of science itself. It is time for science to be understood so broadly that its contributions

can be kept from being used for purposes that could harm mankind. Science should provide more protection from the possible deleterious effects of its increasingly powerful products. Scientists and technologists should provide fail-safe capabilities for their creations, the ability to anticipate undesirable by-products, and leadership in the elimination of those problems of the world for which man has been responsible.

This may seem naive and impractical to some. However, continuation of the present exploitation of science and technology is truly naive.

Now, why is it that we speak of *new* sciences and of *new* engineering areas? We are aware that most of the activities we have discussed are subsumed under existing disciplines such as linguistics, information theory, systems science and other current activities of scientific and technological groups seeking to unite their efforts to promote more general studies of these subjects.

We speak about them as new sciences because information has dynamic properties. That the pen is mightier than the sword is an old manifestation of the power of information. Yet relatively little effort has been made to really understand information dynamics in terms of our present needs. It is hopeful and promising to see societies for cybernetics, general systems, information science, and so on, becoming concerned with some of the aspects of the problem. However, it is quite probable that the problem is now growing at a rate beyond the capabilities of these professional societies. The dynamics of information, its communication, and its ultimate application in the various systems that exist or that have been created by man are the *dynamics of chain reactions*.

It is time to create a synthesis that can include all uses of information. We must plan for the most effective understanding of the good and the bad effects so that we can begin to cope with the complexity that we are creating with the increasing reactivity of chain reactions.

Out of such new organization of disciplines may come new horizons, new approaches, new concepts and new capabilities that are needed for considered growth and implementation in research and development of all fields. If we consider the field of information science, it is quite probable that a group working to develop information in one discipline may find resources, methods and devices that will be useful in information-seeking in other fields. Approaches and instruments that have been developed in scientific research seldom stay in the narrow field for which they were created. The utility of methods at first used in, say physics, such as spectrography, nuclear magnetic resonance, calculating, measuring of any sort, after their fundamental principles were defined and research performed on the fundamental phenomenon, find application in chemistry, the life sciences, and lead ultimately to quality control of the production of goods and their control of our society.

New principles, points of view, philosophies, definitions and specifications with respect to information will be useful in many, if not all, fields. The exploratory work on information in one field should provide insight, impetus, and working tools for other fields. This should offer lower costs in terms of manpower and money. Of even more value than lowered costs, in times of great need, such as the present, will be a shorter time lapse from initiation to broad use.

Similar statements can be made about communication science, systems science and the science of applications. Understanding of the systems that resulted from systems analysis, beginning with the development of the aerospace field, has led to new approaches in the many fields, including the behavioural and life sciences. On the other hand, as more understanding of the operation and communication of information in the organs of animal organisms is achieved, new insights into social systems will evolve. It is even possible that the fail-safe philosophy developed for the operation of nuclear reactors could be translated into a method of international cooperation. For example, if international cooperation reaches an impasse because of a disagreement over territory, trade agreements, flow of materials or some other national phenomenon, agreement on a fail-safe philosophy

might be useful. Instead of pushing on to the brink of war, when non-agreement occurs, the relationships among particular nations will drop to a lower level of safe interaction from which war is not possible. This will gain time and provide a meeting ground for consideration of all possible alternatives in a more objective atmosphere.

The repertoire of illustrations of the possibilities of use of these sciences is almost endless and will be limited only by the imagination and creativity of the disciplines involved.

Research and development have reached a stage where the unit cost of knowledge and productivity is increasing exponentially. Productivity of science and development is also increasing proportionately. An even greater exponential applies to the need for increased research. Information and knowledge are the by-products of these increases. They, in turn, are increasing exponentially. Each new unit of knowledge brings with it the possibility of interrelationships with other fields. Many of these change those fields so that they, in transmitting their changed states, create new possibilities of interchange in still other fields. So while we are creating more and more knowledge, the by-products of that knowledge are increasing at an exponential rate. This is probably an exponential product of the exponential growths. The ability of mankind to cope with the complexity thus created seems to lag farther and farther behind the creation of new interactions. For that reason, more and even better research is needed and again, this will create ever-increasing cost. We cannot slow down, we *must* increase our research efforts as well as our reflection on its results.

We need to increase the ability to reflect, to think about, and to create more research and at a lowered unit cost. The mass production of research, in fact, the achievement of its automation, depends on new concepts and new directions.

Hopefully, these new approaches will provide a design for coping with the problem of rising complexity.

PAPER 5

**TRENDS AND DEVELOPMENTS IN CHARACTER
AND PATTERN RECOGNITION**

by

L. A. Feidelman

**Auerbach Corporation
Philadelphia, USA**

SUMMARY

Up-to-date developments in the fields of pattern and character recognition are described, and technical and economic possibilities for the near future are forecast by surveying present technical principles, present and potential areas of application, and significant trends.

It is indicated that although character and pattern recognition are based on the same techniques, their areas of application are quite diverse. Character recognition techniques have been incorporated into commercially acceptable computer peripheral equipment in the form of character readers, whereas pattern recognition work is still in the experimental stage.

TRENDS AND DEVELOPMENTS IN CHARACTER AND PATTERN RECOGNITION

L. A. Feidelman

1. INTRODUCTION

We, as human beings, have constantly looked for ways to make life easier, be more productive, and extend our capabilities. Such pursuits have resulted in the design of the computer.

The computer's main function is the control and manipulation of data. Although this function has been carried out in an efficient and expeditious manner, the means of presenting this data to the computer are slow. In most cases, the common technique has been to take data from its source and have a keypunch operator translate it from human-readable to machine-language form. However, such translation has proved to be too costly, time-consuming, and unreliable.

The problem was then to design equipment that would efficiently and economically provide a direct interface between the computer and its data environment. A solution to this problem has been accomplished via the technologies commonly known as pattern and character recognition. Pattern recognition denotes the automatic identification of all patterns. Character recognition specifically relates to the automatic identification of alphanumeric characters and symbol patterns. Character recognition, which is technically included under pattern recognition, warrants special consideration since it is implemented in a commercially available device. The more complex problem of automatic recognition of general patterns is basically in the realm of research.

This paper presents a description of the pattern and character recognition technologies, present and potential application areas, and significant technical and economic trends.

2. DEFINITION OF TECHNIQUES

Pattern recognition can be defined as a technique for automatic identification of a given figure or arrangement which is known to belong to one of a finite set of pattern classes. This figure may relate to a missile launch site on an aerial photograph, a tumor on an X-ray, or resistor and capacitor symbols on an electric circuit diagram. The automatic reading of patterns replaces the present method because it eliminates visual inspection of each film frame. Not only is this task physically exhausting, it is also prone to errors.

Character recognition is a technique for automatic identification of alphanumeric characters or symbols. This technique has relatively clearly defined property characteristics as compared with the general class of patterns. The character recognition technique has been used in a device called a character reader, which is primarily a replacement for the keypunching and card reading operation. The character reader permits printed, type-written, or handwritten data to be entered directly into the data processing system from the source document. In practical operation, this direct conversion is not always possible due to uncontrolled data preparation conditions so a retranscription of data via typing is necessary. However, this typing operation has proven to be faster, more reliable, and more efficient than keypunching; it also requires fewer hours of training.

3. EQUIPMENT PRINCIPLES

The same equipment principles may be employed, with some modifications, in both areas although character recognition is simpler from a recognition stand point. The general configuration for a pattern or character recognizer is shown in Figure 1.

3.1 Transport

The transport unit is used to move the film or paper form past the scanner. Since these units are mechanical in nature, they must position the data to be read and move it at a proper speed. The present units have been perfected to handle various sizes, types, weights, and thicknesses of forms, but they are still the slowest part of the system.

3.2 Scanner

The function of the scanner is to convert the pattern appearing on the film or paper form into some analog or digital representation. There are two basic types of scanners: magnetic and optical. Magnetic scanners, which apply only to character readers, use a magnetic read head to sense variations in magnetic flux produced by the difference between the magnetic mark and its background. Optical scanners, on the other hand, employ a light source to detect contrasts between the pattern and its background.

Different optical scanner types being employed include rotating mechanical discs, flying spot scanners, parallel photocell or "retina" photocell sampling techniques, or vidicon television camera tubes. At present, the flying spot scanner is the most commonly used optical scanner for pattern and character recognition due to its ability to adjust the scan pattern and its high resolution. Within character recognition, the retina photocell arrangement permits the fastest sampling of characters. The vidicon tube scanner, used primarily for character recognition, is inherently the fastest scanning technique, but it can read only a limited number of characters on a form (approximately 45 characters per form) due to the tube resolution.

3.3 Recognition

The recognition unit, which is the heart of the system, has the function of extracting significant properties from the pattern and identifying them according to class. In early character recognition work, this recognition function was implemented in a special-purpose hardwired device. At present, both character and pattern recognition rely basically on a computer or computer-type device, with some programmable control, for the recognition function.

3.3.1 Character Recognition

Property definition for character recognition relates to the formation differences within the given character set to be read. This character set is defined by the font or style of the characters and is determined by whether alphanumerics or numerics only are in the set. Identification of a character is accomplished by matching patterns from the scanner against reference patterns for each character.

The font most widely used in the United States and adopted as a standard by the American Bankers Association is E-13B (see Figure 2), which can be used to represent only 10 numerics and four special symbols. Another font (see Figure 3), developed by Compagnie des Machines BULL - General Electric, is capable of representing all the characters in the alphabet as well as all the numeric symbols; and has been adopted as a standard by the European banking community.

The significant property differences among the characters in the E-13B font (see Figure 2) are defined by the voltage waveform produced by a line-by-line scan. Identification is accomplished by matching against reference waveforms. The BULL font shown in Figure 3

differentiates characters by width variations between each seven-stroke character. A character is identified by comparing sequence and number of narrow and wide gaps with stored codes for each of the alphanumeric characters.

Optical readers fall into three classes: character, mark sense, and bar code.

Optical character readers (OCR) recognize the actual character by directly reading its outline or shape. The present OCR readers are capable of reading a large variety of character fonts that may be printed, typewritten, or hand lettered. The sophistication of devices varies; some read only a single font while others can read multiple fonts. An attempt towards standardizing fonts has produced the USASI (see Figure 4) and ISO-B (see Figure 5) type fonts adopted by the United States and certain European countries, respectively.

The identification of such characters has been implemented basically by a matrix matching technique where the scanned elements of the characters are matched against references by means of resistor matrices. Another prominent technique, called stroke analysis, differentiates characters by the position or frequency of vertical and/or horizontal strokes. The character pattern is then matched against a truth table indicating stroke formations for each reference character. Curve tracing, a newer technique employed for handprinted recognition, follows the character outline indicating certain features such as character splits, line intersections, line magnitudes, and line straightness.

Mark-sense readers sense the physical position or location of marks on a document correlating the mark position to a previously defined equivalent character. This technique requires preprinted forms. Most present OCR readers provide mark-sensing as an option.

Bar-code readers utilize thick line or bar representations of characters. Each character is defined by a given number of long and short bars, and can be identified by matching the code against a character reference table.

3.3.2 Pattern Recognition

Pattern property definition and identification techniques can be described only in general terms since they are highly dependent upon the specific characteristics of the patterns to be recognized. Property definition relates to the separation of patterns into classes or categories. This separation is attempted by determining the relative invariance of patterns in terms of significant, relevant, and interdependent features and their changes in the time domain. Such pattern features include size, location on film or paper, curves, slopes, symmetry, and grey levels.

The recognition techniques must identify the specific pattern in terms of a unique combination of pattern features which have been changed due to noise. The techniques used are quite numerous with no one technique gaining complete general acceptance.

A significant amount of research has been devoted to the development of "learning machines" involving continual adjustment of the recognition logic to new combinations of patterns or new probabilities of a pattern's occurrence. These techniques basically use a statistical approach combined with property weighting schemes for identification. Another fundamental technique is to construct a decision tree with a node relating to one property. The specific pattern is identified by a process of elimination. Curve tracing, as employed for character recognition, is a third technique; and many others involving topographical considerations are presently being considered.

4. PRESENT AND POTENTIAL APPLICATION AREAS

The application areas for character and pattern recognition are diverse. Character recognition work is business-oriented, and directed toward supplying data more efficiently

to the computer. Pattern recognition is science-oriented, and directed toward the analysis of visual information which has little semantic value.

4.1 Pattern Recognition

Present pattern recognition has been concentrated basically in three areas:

(a) *Aerial Photography*

The enormous amount of aerial photographic interpretation now required has resulted in a definite need for automatic identification of tactical and strategic targets and discrimination of terrains as a preselected aid to photo interpretation by humans. The problem of determining a unique representation of a target and accounting for its variations is still not solved. Also, extraction of noise from pattern is a serious problem. The recognition of terrain types is somewhat easier since analysis of picture detail and grey level classification is a solution key. Once the properties are defined, then some statistical approach, such as Bayes procedure, can be employed.

(b) *Medical Field*

Pattern recognition as an aid to medical analysis is concerned with such problems as (a) analysis of X-ray films for tumors, (b) irregularities in blood cells or other parts of the body, and (c) classification of blood cells. Again the basic problem is defining the property. Most work is accomplished by sampling results and looking for significant characteristic differences rather than using any preset decision rule.

The objective, of course, is to relieve the doctor from examining all possible data when he need be concerned with only selected groupings. This change would reduce the valuable time a doctor must spend looking at X-rays, for example; it also reduces the eye fatigue that may lead to an improper analysis.

(c) *Voice Prints*

Voice-print identification is a method by which people can be identified from a spectrographic examination of their voice. This method, which is analogous to fingerprint identification, is to be accomplished by uniquely defining people according to their utterances. The technique is based on examining the amplitude contours as affected by people's vocal cavities and articulators. The size of vocal cavities and articular uses for different people are claimed to be unique. Speech contours of various people are still being examined to determine (a) complete uniqueness, (b) voice changes of people with time, and (c) disguising of voices.

4.2 Character Recognition

OCR devices are slowly taking over the market from magnetic character readers since the OCR offers high flexibility. It can read various types of source data, offers increased reliability, and is substantially competitive in cost with the magnetic reader when single fonts are used. New OCR devices incorporate the mark-sense and bar-code features. The magnetic character reader's basic advantages are as follows: (1) the security problem is better handled because magnetic characters cannot be forged as easily as regular recorded characters; (2) it can read when dirt is a problem; and (3) it can read whenever over-stamping is a problem.

The present purchase price of commercial magnetic readers averages around \$80,000. The prices for optical readers range from \$90,000 to \$600,000 depending on the speed and sophistication of recognition cost (rentals run from \$3000 to \$20,000 a month).

4.2.1 The character reader application has fallen into three basic areas:

(a) *In-House Applications*

In this situation the character reader is replacing the keypunch machine. All document preparation and character reading are done in a centralized location, permitting tight document control. If the present application requires approximately eight to ten keypunch operators, then it becomes a definite candidate for the use of a character reader. In this situation, the single font reader is recommended due to the controlled conditions.

(b) *Turn-Around Documents*

These documents are prepared by the computer, sent out, returned, and then read back into the system. The single font reader is excellent for this situation.

(c) *Field Documents*

This represents the most uncontrolled situation, and the character readers are becoming more adept at reading different forms with different kinds of typing; they are even going into the handwritten application. The bulk of present equipment (multifont and handprinted readers) is concentrated on meeting applications in this area.

4.2.2 A representative list of uses for character readers is as follows:

(a) *Oil Industry*

The oil industry has been using OCR devices to read embossed cards where the customer has a credit card and the data are transferred from the credit card onto the document by means of imprinters. Added to this is the cost of the particular purchase which can also be put in by the imprinter, by mark sense, or by hand in the future. The oil industry also has turn-around document situations in which statements are sent out to customers and have to be returned and read by the system. Invoice billing of the various service stations is another area where the oil companies can use character readers. Small-cost single readers are applicable. Handprinted character readers can be used although they are not necessary.

(b) *Transportation*

The airlines in the United States have taken an extensive look at optical readers. TWA and United are using optical readers to read the preprinted ticket numbers at the bottom of the ticket. A future use may be to have the reader read tickets and issue boarding passes or read ticket baggage. This application will require the character reader to go from its normal batch processing jobs to real-time jobs with the physical separation of the scanner unit from the recognition unit. Data can be transmitted by means of facsimile transmission between the scanner and the recognition unit.

(c) *Banks*

The banks have used magnetic character readers for reading preprinted check information. However, a trend towards OCR devices is apparent in banks; they are using them to read name and address changes, installment loan information, stock transfer information, etc. for input to the computer.

(d) *Inventory Control*

One problem of inventory control is getting the source data from various distribution points into a centralized data processing system where they can be processed by the computer system. Input is at present accomplished by keypunching. Conversion to source data automation equipment such as OCR or magnetic encoders is at present being performed with success and should continue.

(e) *Insurance Company*

The insurance companies have been using the optical readers, basically mark sense, to read application forms and have not fully exhausted all the uses of the reader.

(f) *Post Office*

The US Post Office is now experimenting with multifont alphanumeric readers to read typewritten business mail (which accounts for approximately 70% of the total). The readers are capable of reading the entire alphanumeric address and zip code. Future developments include a character reader capable of reading handwriting and printing.

(g) *Publishing Houses*

Optical readers are used to convert subscription forms and returned customer statements into machine language. A new potential application is the reading of book covers returned to the publisher for money refund.

(h) *Social Security*

The US Social Security Office in Maryland is utilizing a multifont character reader for directly reading employee tax information as received by employers.

5. TRENDS

Awareness of both character and pattern recognition has increased significantly in the past year basically due to the commercial acceptance of the character reader. Once considered a research device, this reader is now taking its place in the data processing system. Therefore, pattern recognition, which has taken secondary consideration, can now be given more technical attention. The future challenges to the scientist and engineer definitely lie in the area of pattern recognition.

While pattern recognition work is involved in property definition and determination of proper identification algorithms, character recognition work is based on increasing the reader flexibility, reliability, and speed while reducing the costs to cover a wider area.

In this respect there are eight basic trends in character readers.

(a) *Software*

There is an upward trend to use a programmable unit for recognition rather than to rely on special purpose hardware. Programmable units increase flexibility; e.g., they can read forms having different character fonts and field formats. In addition, the programmable unit can be used for data extraction, sequencing, and manipulation to reduce the computer load further.

(b) *Recognition of Handwriting*

The work being done on the recognition of handwritten characters can be divided into two classes: handprinted and script. The ability to read numeric handwritten characters already exists in the readers manufactured by IBM and Optical Scanning Corporation. Recognition Equipment has recently announced the capability to read handprinted alphanumerics. The IBM 1287 Optical Reader, for example, can read handprinted numeric digits and five alphabetic control symbols, but a glimpse at the rigid set of rules shown in Figure 6 emphasizes that the concept is still quite restricted in practice. However, reading of script characters is only in the developmental stage, the Post Office being the primary customer.

(c) *Context Recognition*

Context recognition, a long-range effort to reduce reject and error rates, is an attempt to simulate the human ability to apply contextual significance to characters or elements which might otherwise be devoid of meaning. When a person reads, the legibility of individual letters, or even of individual words, is not usually

critical. Human beings "read" or perceive letters within the context of the entire word, and words within the context of an entire sentence. Consequently, the reader easily identifies "Sxrxt" (where the "x"s represent garble), in the phrase "2231 South 12th Sxrxt", as "Street" even though only 50% of the letters are readable. This identification is possible because of the conditional dependency of letters and words in human communication.

Although context recognition is not yet sophisticated enough to become a major factor in a recognition scheme, it can be used as a back-up method for identifying illegible characters. The most obvious advantage of this technique lies in its potential to identify a complete word in which one or more characters may present serious recognition difficulties.

(d) *Off-Line Versus On-Line*

One of the benefits of OCR is placing the input peripheral device off-line (in most cases). The on-line card reader has slowed down the computer and has, in some cases, required a separate computer for preprocessing. The trend of character readers is toward independent off-line units producing magnetic tapes for computer processing.

(e) *Speed*

Another, but less critical, area of developmental emphasis in character readers is in speed. Reading speed is at present limited by the amount of time required to mechanically move the document past the reading station. The overlapping of the reading and transport operations is accomplished by using storage tubes (i.e., vidicon scanners) or by reading "on the fly". Speed can also be increased by using form controls which perform selective field reading, and skip blank spaces. The reliability of the character reader not only affects its accuracy, but also has a significant impact on document-reading capability; actual reading speed is obviously affected if a document must be read more than once.

(f) *Improvements in Reliability*

Naturally, reliability in the form of low error and reject rates is a prime consideration in all the development work being done on character readers. One approach to reduce these rates is to improve the resolution of the scanning units and thereby increase the number of sample points from which the equipment can make an identification. As previously mentioned, Philco-Ford Corporation is using a cathode-ray tube that has a resolution of 2000 optical lines. Even better resolution can be expected in the near future.

The reading reliability of the character readers, in terms of reject and error rates, has improved substantially due more to the source document preparation control, typist training, proofreading, and special checks within the character reader than to the recognition logic itself. The present reject rate for in-house form preparation is presently two to three percent. Based on improvements in preparation, this reject rate will drop significantly to below 1% in the next five years, and error rates will fall to below 0.5%. Key punching error rates are presently 1.5%.

(g) *Cost*

Present commercially available character readers are designed for large-scale operations (more than 10,000 documents per day), in which cost can be justified. There is, however, a definite need for a low-cost single-font character reader (approximately \$20,000) which could read fixed-format single-document types. In view of the recent character set standardization, it would appear that a trend toward such a device is now likely.

(h) *Remote Scanners*

The use of remote scanners connected in a time-sharing configuration with a centralized recognition unit is within the state of the art and can be expected soon.

6. CONCLUSION

Our world exhibits the confidence that no scientific challenge is too great to meet sooner or later. Pattern and character recognition development is no exception, and present work indicates that the next decade will see amazing results. The character reader has proven its ability to meet such a challenge; pattern recognition is a step away.

REFERENCES

1. - *Character Recognition.* London, British Computer Society.
2. Feidelman, L. *A Survey of the Character Recognition Field.* Datamation, February 1966.
3. Feidelman, L.,
Katz, J. *Scanning the Optical Scanners.* Data Process. Mag., October 1967.
4. Feidelman, L.,
Katz, J. *Auerbach Special Report on Optical Character Recognition.* Auerbach Data Handling Reports, March 1967, Auerbach Corporation, USA.
5. Fischer, G.L., Jr.
Pollack, D.K.
Poddock, B.
Stevens, M.E. *Optical Character Recognition.* USA, Spartan Books, 1962.
6. Kanal, L.N.,
Hamblar, K.K. *On the Application of Discriminant Analysis to Identification in Aerial Photography.* Proc. 7th Nat. Conv. on Milit. Electron., September 1963, USA.
7. Kersta, L.B. *Voiceprint Identification.* Nature, London, Vol. 196, 29th December 1962, pp.1253-1257.
8. Rosenfield, A. *Automatic Recognition Techniques Applicable to High - Information Pictorial Inputs.* I.R.E. int. Conv. Rec., Vol.10, 1962.
9. Stevens, M.E. *Automatic Character Recognition - State of the Art Report.* USA Department of Commerce, May 1961 (PB 1616 13).
10. Unger, S.H. *Pattern Detection and Recognition.* Proc. Inst. Radio Engrs, Vol.47, October 1959, pp.1737-1752.

The discussion on this paper follows on p. 61.

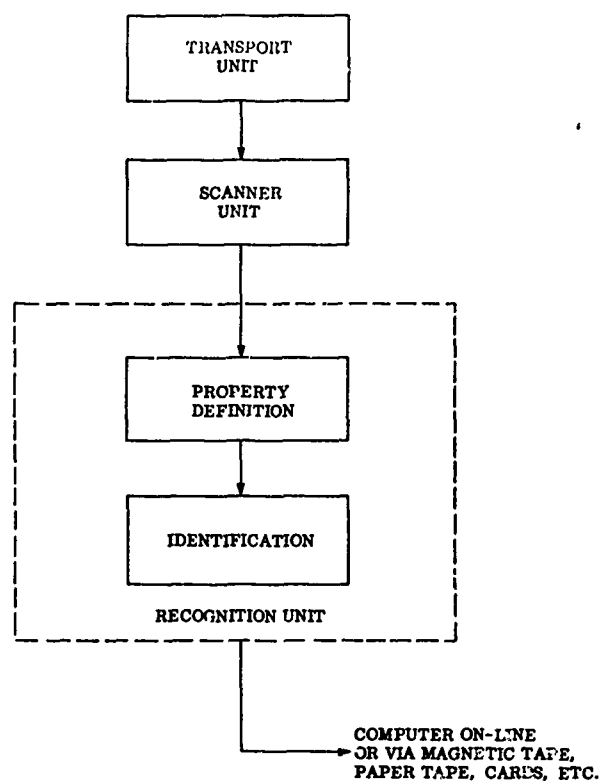


Fig. 1 Block diagram of pattern/character recognizer reader

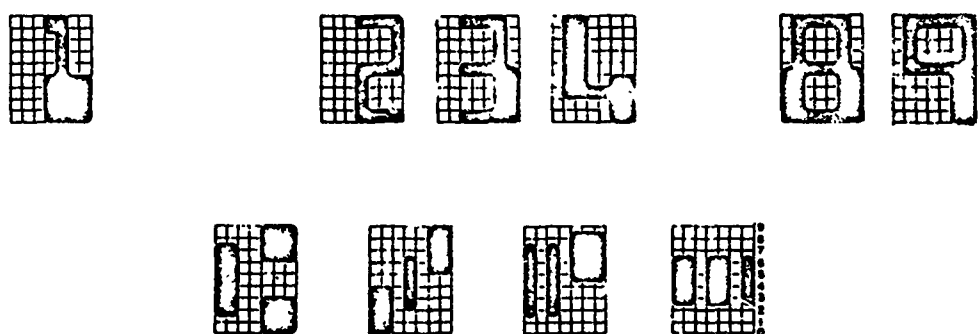


Fig. 2 Sample of E-13B font characters

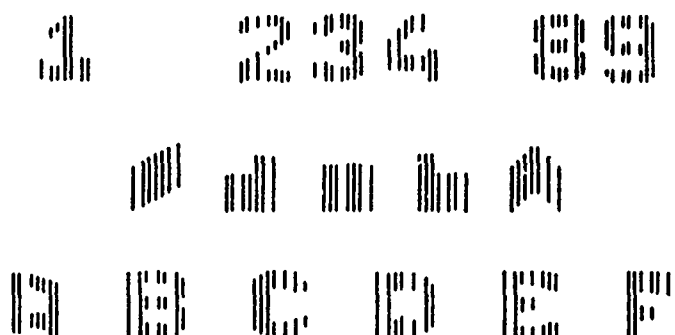


Fig. 3 Sample of BULL magnetic reader type font characters

A B C D E F G H I J K L M
 N O P Q R S T U V W X Y Z
 0 1 2 3 4 5 6 7 8 9
 . , : ; = + / \$ * ^ & |
 ' - { } % ? [] ^ _
 Ü Ñ Ä Ö Å Æ £ ¥

Fig. 4 USASI font

A B C D E F G H
 I J K L M N O P
 Q R S T U V W X
 Y Z * + , - . /
 0 1 2 3 4 5 6 7
 8 9

Fig. 5 ISO Class B font

RULE	CORRECT	INCORRECT
1. WRITE BIG	0 2 8 3 4	0 2 3 3 4
2. CLOSE LOOPS	0 6 8 8 9	0 6 8 8 9
3. SIMPLE SHAPES	0 2 3 7 5	0 2 3 7 5
4. DO NOT LINK CHARACTERS	0 0 8 8 1	0 0 8 8 7
5. CONNECT LINES	4 5 1	4 5 1
6. BLOCK PRINT	C I S T X Z	C I S T X Z

Fig. 6 Handwriting rules for IBM 1287

DISCUSSION

R.J.Dubon: Are there any plans, in the USA, to standardize the fonts used in the printing industry?

L.A.Feidelman: Standardization of type fonts, paper and ink is required to aid character recognition. A standard has been drawn up by representatives of US computer manufacturers. There is also an International Standard Font.

P.Molzberger: Present-day multifont reading machines require expensive hardware or very long programs additional to that of the main computer. Is there likely to be a trend in computer manufacturing to make it possible to run an ordinary computer as a highly specialised, parallel working recognition-logic complex?

L.A.Feidelman: In a parallel working system it would still be necessary to have a separate processor, separate memory, etc. so that this is probably not the answer. In many ways an off-line recognition system is the better method.

H.F.Vessey: Why have some organisations ceased using optical scanning for bibliographic input to the computer?

M.S.Day: NASA has tried using optical readers for input, and although it works, it has been found to be too costly at present. Hopefully, costs will come down in the future.

S.Skounal: Can you foresee the use of central recognition hardware providing input to a remote linked computer?

L.A.Feidelman: This is a possibility. Airlines plan to have remote scanners at boarding points and information from passenger tickets will be passed to a central facility to see if the information matches that already held before the passenger is allowed aboard.

PAPER 6

EFFICIENT TRANSFER OF
TEXTUAL INFORMATION

by

J. W. Altman

American Institute for Research,
Pittsburgh, USA

SUMMARY

Three problems which arise in attempts to achieve efficient provision of textual information to scientists and engineers are defined:

- (a) Text-sensitive tasks of scientists and engineers have not been delineated and analysed sufficiently to define clearly their requirements for textual information support.
- (b) Methods have not yet been established to permit characterisation of text in terms which support the development of a technology for efficient transfer of text to users.
- (c) Little concentrated effort has been put into attempts to establish lawful relationships between text - sensitive tasks and characteristics of text.

Ways of tackling these problems are suggested.

EFFICIENT TRANSFER OF TEXTUAL INFORMATION

J.W. Altman

1. THE PROBLEM

Price⁹ has described the growth of scientifically based knowledge in graphic terms:

"... any young scientist, starting now and looking back at the end of his career upon a normal life span, will find that 80 to 90 percent of all scientific work achieved by the end of the period will have taken place before his eyes, and that only 10 to 20 percent will antedate his experience (pp.2-3)."

Under such circumstances, the need for efficient transfer of information from authors to potential users hardly requires belabouring. Yet, three central problems remain:

- (a) Text-sensitive tasks of scientists and engineers have not been delineated and analyzed in such a way as to define clearly their requirements for textual information support.
- (b) Methods have not yet been established which permit characterization of text in terms which support generation of a technology for efficient transfer of text to users.
- (c) Little concentrated effort has gone into attempts to establish lawful relationships between text-sensitive tasks and textual characteristics.

It is the purpose of this paper to review some of the findings already available concerning these issues and to suggest ways in which they might be resolved.

2. DEFINITIONS

The use of the terms "transfer," "text," "information," and "efficiency" in this paper requires some definition. "Transfer" is used here to refer to the process of transmitting knowledge from an author to potential users of that knowledge. It is assumed that authors will, in the main, follow the precepts commonly accepted for effective technical writing. It is also assumed that any machine or manual processing of the author's text that is implied here can be implemented within the general scope of available techniques. Consequently, this paper will not deal with textual processing, storage, and retrieval as such. Rather, it will emphasize how the requirements for such processing, storage, and retrieval should be established. "Text" is used here to refer to written narrative, tables, illustrations, graphs, formulas, or any combination of them. "Information" is any identifiable influence on behaviour other than a direct physical restraint or physiological impairment. Textual information or text-carried information is thus behaviour which can be demonstrated to be a result of exposure to given text. For present purposes, it can be seen that information is in the behaviour of the user rather than being a directly measurable characteristic of text.

"Efficiency" will be discussed primarily in terms of achieving quality scientific and engineering task performance with minimum expenditure of user time to exploit the supporting text.

3. PHASES OF TEXT USE

My primary structure will reflect the principal phases of textual use, as follows:

- (a) Screening.
- (b) Gaining and maintaining awareness of a scientific or technical field.
- (c) Application of text-mediated information.

Figure 1 shows a schematic summary of relationships among these three phases and their major sub-phases. I will not tarry here to discuss these phases of text use but will discuss their salient features in passing as I attempt to identify some of the more critical information transfer problems for each phase.

Perhaps a word of caution would be in order here, however. The phases of text use presented here are somewhat arbitrarily chosen to support the current discussion. It is not my intention to imply a reflection of a well-mapped domain of human behaviour, for such mapping has not yet taken place.

3.1 Screening

Document screening can be conceived as occurring in three principal waves. The first wave involves selection of a set of candidate documents from the total body of scientific and technical text. The second wave involves selection from the set of initially considered documents those that will receive serious scrutiny. The third wave includes reading of documents and selection of those parts that will be remembered or applied. This third wave is so inextricably entwined with maintaining awareness and application that I shall deal with only the first two waves here.

3.1.1 Search

(a) *Retrieval Based on Recall*

The sequence from identification of a technical task to be performed to the availability of a tentative bibliography from which to select documents for detailed review is perhaps the phase currently containing the most dysfunctions between textual information systems and scientist-engineer needs. It is hardly surprising that this is a phase of text use beset with difficulty since it is one in which it is necessary to go from the entire body of available scientific and technical documentation to a relatively short list of reasonable candidates. Neither is it surprising that there has been a preoccupation, over the last decade or so, with retrieval schemes to support the purposes of this phase.

I shall not attempt hereto review the current status of document retrieval systems. Rather, I would like simply to draw a distinction between recall systems and non-recall systems and to discuss some of the salient features of each. Recall systems are based on the user's recollection of a specific document. I shall not discuss the mechanics of retrieving specific documents, whether remembered with fidelity or semi-reliably. However, it may be appropriate to point out that the citation index is one technique for projecting one's memory ahead. That is, if one remembers a given document as being relevant to a task at hand, by reference to a citation index, it is possible to identify subsequent documents which cited the remembered document - which presumably are likely to be related to the remembered document.

Recall-based retrieval tends to follow rather classic patterns and to be accomplished without undue difficulty. Whenever one either remembers specific documents or knows who is doing the most related work, retrieval of appropriate documentation is likely not to pose serious conceptual problems. Indeed, "in" members of "invisible colleges"⁹ generally claim little difficulty in maintaining currency with the scientific-technical documentation of greatest relevance^{3, 4, 5}.

(b) *Retrieval Based on Descriptors*

If a researcher is not willing to depend upon his own memory or the knowledge of his friends to ensure comprehensive coverage of relevant documents, he is faced with a much more difficult problem. True, he may have a whole armoury of descriptor languages, structural models, organizing principles for files and codes, search procedures, and automated document retrieval aids¹¹; but it is impossible to pretend that the sum total of our existing retrieval systems will necessarily suffice to insure an adequate search of all the appropriate documentation. The inadequacies of existing retrieval systems are likely to become particularly apparent in coping with the newer and less rigorously delineated fields--precisely those in which most of the exciting action is likely to be taking place.

Since the body of reasonably current scientific-technical literature is too large for any of us to thumb through in a lifetime and is growing at an accelerating rate that would put us further behind at the end than when we started, we must depend for document retrieval upon procedures that will preclude our having to look at any significant portion of the total body of scientific-technical text. If we are to avoid the vagaries of random sampling, this means that we must depend upon analogs of analogs. Each scientific-technical document represents a symbolic analog of some phenomenological domain. Yet, the documents themselves represent too bulky a phenomenological field to be dealt with directly. If we are to have facile ways of dealing with this body of symbolic analogs, we must depend upon yet greater abstractions (simplifications) of these analogs to the real world. But what kinds of analogs will suffice for this purpose?

I cannot hope to review here the great number and variety of schemes which have been devised to provide a facile analog to a corpus of documents. Instead, I will explore briefly one approach which seems to have promise for clarifying the underlying logic of document retrieval systems as well as practical implications for improvement in retrieval, particularly automatic, of documents.

Ossorio⁶ has demonstrated the possibility of classifying documents in n-dimensional Euclidean space. Without espousing Ossorio's particular definition or approach, it is possible to see how dimensionalizing the domain of documents and the phenomena to which they refer can enhance retrieval of documents. Such dimensionalizing can be based on multivariate statistical analysis as was Ossorio's or on more direct analysis and representation of the phenomenological fields with which science and technology deal.

Let us suppose that we have structured a given field according to n orthogonal (uncorrelated) dimensions. Let us further imagine that we are able to scale each of the descriptors for any document according to one or more of the underlying dimensions which provide gross structure for the field. Finally, let us assume that we can frame our request for a document search in descriptors, each of which can also be scaled according to one or more of the underlying dimensions.

Given these assumptions, it can readily be seen that, even though different terms may be used to describe documents and requests for document search, documents and search have a common frame of reference in the underlying dimensions used to structure the phenomenological field. With such a common frame of reference, it should be possible to derive a meaningful quantitative basis for matching documents and requests.

It may be an adequate first-approximation assumption that the probable utility of a document will be a monotonic increasing function of the following:

$$U = \sum_{d=1}^D \sum_{t=1}^t \left(1 - \frac{d}{D}\right)^n$$

where:

U is a figure of merit for probable utility of a given document for a given request,

\sum^D indicates a summation across all of the dimensions used to structure the field,

\sum^t indicates a summation across all of the terms scalable on a given dimension (given n terms for the document scalable on the dimension and k terms for the request so scalable, there would be nk factors under \sum^t).

d is the dimensional scaling difference (without regard to sign) of a given term in the document description versus a given term in the retrieval request,

D is the range of a given dimensional scale, and

n is a power function probably best determined by empirical study of the rate at which utility tends to fall off as a function of distance between document and request along a given dimension.

The probable utility of a given document is not only a function of its convergence with a potential user's needs, but may also be influenced by the redundancy of that document with another. That is, the user is likely to have use for a given piece of information only once. If it is contained in more than one document, all but one document may be superfluous. Consequently, it may be desirable to present to the possible user only the most recent document of any sets which are extremely close on all dimensions.

3.1.2 Bibliographic Review

The initial search for documents relevant to a given intended use should result in a bibliographic listing of some type. It has been observed in a number of situations^{2, 10} that the accuracy of initial document screening is not greatly influenced by the extent of textual detail available to support this screening. Also, such screening with highly attenuated text can be accomplished generally with about 30 to 50 percent saving in time over the use of full text. Consequently, it seems entirely appropriate to depend upon highly attenuated text for such initial document screening.

Descriptive titles should suffice for short, relatively homogeneous articles. Brief indicative abstracts or topic descriptors should suffice to represent longer or more varied documents for purposes of initial screening by the user.

The merits of attenuated representations of documents for initial screening by a potential user suggest that automatic retrieval systems which result in a bibliographic output modestly amplified by key descriptors may, indeed, be compatible with user needs and behavioral tendencies. Of course, this does not mitigate the need for such retrieval systems to furnish appropriate bibliographic output for review.

3.2 Gaining and Maintaining Awareness

Whether one is retrieving documents for immediate application or simply accumulating information for possible future purposes, it is required of the user that he become aware of textual content through reading. This truism need not hide the fact, however, that the textual requirements for maintaining current awareness and gaining awareness of text for an immediate and specific purpose are different. I will discuss the textual requirements of maintaining general current awareness of a field prior to discussing textual needs for specific purposes.

3.2.1 Maintaining Current Awareness

Several studies^{2, 7, 8} have demonstrated the utility of attenuated text (abstracts and extracts) in comprehending the essential meaning of scientific and technical articles.

The upper limits for reduction without significant loss of general comprehension tend to be about 50 percent (or up to 75 percent for exceptionally verbose material) for words, 25 to 30 percent for figures and formulas. Such textual reductions result in savings in the neighbourhood of 15 to 35 percent in reading and comprehension time. Attenuation of text to the neighbourhood of 85 to 95 percent of the original can be achieved with loss in comprehension (measured in terms of ability to answer salient questions derived from the full text) from 10 to 50 percent.

Differences between comprehension from full text versus comprehension from attenuated text tend to be maximal for text of intermediate complexity. That is, comprehension levels from attenuated and full text tend to be most similar where comprehension from full text is either very poor or very good.

Before passing on from matters of maintaining general awareness of a given field to more specific applications of text, it is perhaps appropriate to say a few words about possible improvements in selective dissemination for this purpose. In general, the comments made concerning the screening of a body of documents for retrieval of selected documents to serve a specific purpose are also relevant here. Just as specific requests can, at least theoretically, be scaled according to some set of underlying dimensions which structure the field, so also might an individual's general interests be similarly scaled. Documentation having similar multidimensional scaling might then be brought to his attention on a regular basis.

3.2.2 *Gaining Specific Awareness*

If the purpose is to exploit a given set of pre-screened documents for a specific application, the textual requirements are substantially different from those for maintaining current general awareness. Abstracts and extracts can usefully support maintenance of general awareness with interesting savings in the bulk of material disseminated and in reading time, and with only modest loss in comprehension of the major content of the original document. But specific applications usually require greater detail than is provided by abstracts or extracts.

Payne and Hale⁷ found an average loss of about 30 percent (across a number of scientific and technical fields) in efforts to use extensive descriptive abstracts as a source of specific facts. Interestingly, neither the extensive abstracts nor briefer abstracts resulted in a time saving over the use of full text for fact retrieval.

This relative inadequacy of abstracts and extracts for specific applications makes it tempting to conclude that the original document is required in most cases for such purposes. I will present some notions in the following section, however, which suggest that such a conclusion may be premature.

3.3 *Application*

In this section are first presented some of the considerations relating to textual analysis. Then, text characteristics are related to levels of application.

3.3.1 *Textual Analysis*

Boldovici and Altman¹ found that it was possible to break scientific and technical text into elements which they called "textual units."

The operations involved in reducing a technical or scientific article into textual units may be described as follows: (1) the article is first divided into its gross and clearly separable parts; (2) each resultant section is then sub-divided into parts which can be separated without violence to the apparent intent of the author; and (3) the parts are then sub-divided into progressively smaller parts to the point at which it appears that further division would obscure or violate the intent of the author.

The uniformity of textual units resulting from the procedure described above may be tested and improved by applying the following criteria:

- (a) A textual unit is a segment of technical text which may impart information (i.e., reduce uncertainty) when taken by itself.
- (b) A textual unit expresses a complete thought, and can be taken out of context without completely losing meaning.
- (c) A unit of technical text is a segment of written and/or pictorial material which, if further divided, will lose meaning unless reconfigured into essentially its original form.
- (d) All of the material in a textual unit is so interrelated that it would take more words to explain any sub-division than are contained in the original unit.

Given an array of textual units which corresponds to a whole article, textual analysis proceeds by categorizing each unit as to whether it was of primary, secondary, or tertiary importance to the intent of the author's communication.

Textual units judged to bear primary importance to the author's communicative intent are then arranged in a convenient spatial configuration (e.g., from left to right on a page) which reflects the apparent temporal progression of the author's writing, beginning with problem statements and ending with conclusions. Secondary textual units are then attached by lines to the primary units to which they appear most relevant. Tertiary units are similarly attached to secondary ones.

Major types of textual units are defined in Table 1. Idealized configurations of units for deductive, empirical research, and developmental studies are presented in Figures 2-4, respectively.

3.3.2 Levels of Application

We might now use the foregoing characterization of text as a basis for relating it to levels of application, arbitrarily limiting levels to the following four:

- Reflecting simple awareness.
- Reflecting sophisticated awareness and/or collation.
- Reflecting analysis and synthesis.
- Reflecting evaluation.

Reflection of simple awareness may require familiarity with and comprehension of only a limited set of textual units--most likely those involving conclusions and implications. More sophisticated awareness and collation of material from different documents or portions of the same document imply a need for at least selective familiarity with results, interpretations, and proof. Analysis or synthesis of results in terms other than those of the original author implies thorough understanding of the conditions under which the results were obtained and the axioms or rationales which underlay their generation. To evaluate the adequacy of previous work implies a detailed understanding of the derivations and methods as well as the results, conclusions, and implications.

It can be readily seen that there is a hierarchy of applications which has a parallel range of needs for different numbers and types of textual units for minimal appropriate support. That non-chance relationships exist between the nature of applications and needs for different textual units seems beyond a reasonable doubt. However, the nature, strength, and stability of such relationships are essentially unknown at present.

Any of the more extensive or complex scientific and technical documents is likely to have a number of more or less independent logical strings of textual units. Any given

application is likely to have relevance to only a selected sub-set of the entire set of strings in a given document. Given highly accessible and facile information systems in the future, it is not inconceivable that a prospective user would be selectively exposed to only those strings of textual units having likelihood of being relevant to a particular application and in an order most consistent with his scientific and technical task--with subsequent units in the string being presented on demand.

4. CONCLUSIONS

Based on the various findings and arguments advanced thus far, I would suggest the following conclusions:

1. In the preparation of original text:
 - (a) Some of the extensive attenuations which have been accomplished in text without measurable loss of information suggest that there may be considerable reduction yet possible in the volume of text generated through more rigorous attention to the preparation and monitoring of publication standards. This prospect seems particularly rich when it is recalled that the attenuations were accomplished on articles in published journals, which tend to be among the most terse of scientific and technical text.
 - (b) The traditional ordering of text to parallel the logical sequence followed by the investigator may be less than optimum for the individual who uses text. More optimal orders of presentation for use should be sought.
 - (c) The delineation of textual units in the process of preparing text for storage and retrieval may serve as a powerful aid to efficient textual transfer. Its potentials should be investigated.
2. The structuring of phenomenological fields by the methods of multivariate analysis and related rational processes should be explored as an aid to:
 - (a) Retrieval of documents on the basis of specific queries.
 - (b) Selective dissemination of document lists to match individual interests.
 - (c) Characterization and retrieval of individual textual units as well as whole documents.
3. Abstracts and extracts show promise as efficient aids to maintaining current awareness of a field. Only a rudimentary technology now exists concerning this issue. That technology should be strengthened by greatly expanded empirical studies of the characteristics of attenuated text which make it a suitable substitute for full text.
4. Information systems to date have concentrated almost exclusively on the retrieval of whole documents. Work should be initiated on determining the feasibility and value of retrieving individual textual units and strings of related units.

REFERENCES

1. Boldovici, J.A.,
Altman, J.W.,
Toward Improved Techniques for Abstracting and Extracting Scientific and Technical Literature. Griffiss Air Force Base, N.Y.: Rome Air Development Center, January 1967. (RADC-TR-66-674)
2. Boldovici, J.A.,
Payne, D.,
McGill, D.W.
Evaluation of Machine-Produced Abstracts. Griffiss Air Force Base, N.Y.: Rome Air Development Center, May 1966. (RADC-TR-66-150)
3. Garvey, W.D.,
Griffith, B.C.,
Reports of the American Psychological Association's Project on Scientific Information Exchange in Psychology. Washington, D.C.; American Psychological Association, 1965.
4. Herner, S.,
Information-Gathering Habits of Workers in Pure and Applied Science. Ind. Engng Chem. Vol.46, 1954.
5. Menzel, H.,
Information Needs and Uses in Science and Technology. In C. Cuadre (Ed.), *Annual Review of Information Science and Technology.* New York: Wiley, 1966.
6. Ossorio, P.G.,
Classification Space Analysis. Griffiss Air Force Base, N.Y.: Rome Air Development Center, October 1964. (RADC-TDR-64-287)
7. Payne, D.,
Hale, J.F.,
Automatic Abstracting Evaluation Support. Griffiss Air Force Base, N.Y.: Rome Air Development Center, February 1964. (RADC-TDR-64-30)
8. Payne, D.,
Munger, S.J.,
Altman, J.W.,
A Textual Abstracting Technique: A Preliminary Development and Evaluation Support. Pittsburgh: American Institute for Research, August 1962. (RADC-TDR-62-372)
9. Price, D.J. de
Solla.
Little Science, Big Science. New York: Columbia University Press, 1963.
10. Rath, G.J.,
Resnick, A.,
Savage, T.R.
Comparisons of Four Types of Lexical Indicators of Content. Am. Docum. Vol.12, 1961, pp.126-130.
11. Vickery, B.C.,
On Retrieval System Theory. 2nd ed., Washington, D.C.: Butterworth, Inc., 1965.

The discussion on this paper follows on p.76.

TABLE 1

Definitions of Major Types of Textual Units

TOPIC--A statement whose main purpose is to delineate the nature and limits of subordinate content.

PROBLEM--A statement of conditions which justify the establishment of a technical objective.

PURPOSE--A statement of intended technical accomplishment.

DEFINITION--A statement having as its primary purpose the delineation of standard reference for terms used elsewhere in the text.

METHOD--A description of the activities of the investigator or his surrogates.

RATIONALE--Justification of a method.

CONDITION--A description of environmental characteristics presumed to have relevance to some result.

CONSTRAINT--A statement of conditions emphasizing the limits within which the technical effort took place.

AXIOM--A proposition accepted on its intrinsic merit.

LEMMA--An auxiliary proposition accepted as true for use in demonstration of another proposition.

THEOREM--A proposition subject to logical proof.

HYPOTHESIS--A proposition subject to empirical demonstration.

DERIVATION--An intermediate stage of logical manipulation between a theorem and final demonstration or proof.

RESULT--A statement of observations presumed to have relevance to a theorem, hypothesis, or condition of interest.

ENTITY--Description of a technique or device resulting from developmental effort.

COROLLARY--An immediate derivation from a proven proposition.

PROOF--Logical demonstration of the truth of a theorem, within the limits of truth of its related axioms.

INTERPRETATION--The bringing to bear of additional data, logical argument, or relationships to clarify results.

CONCLUSION--A statement of belief about the reality, or reliability of a finding.

IMPLICATION--A statement of a belief about the breadth, depth, or nature of application for a finding.

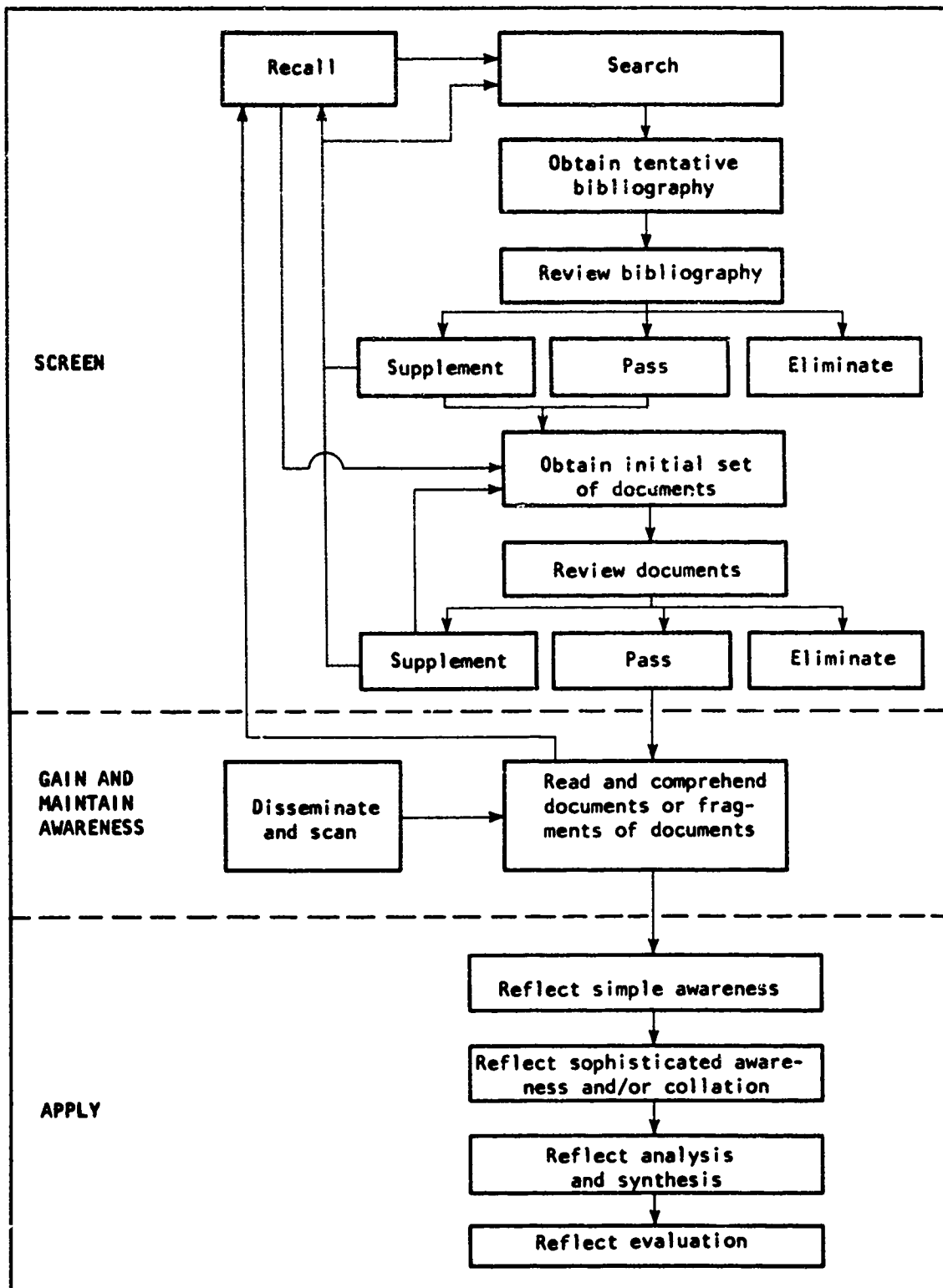


Fig.1 Principle phases in textual use

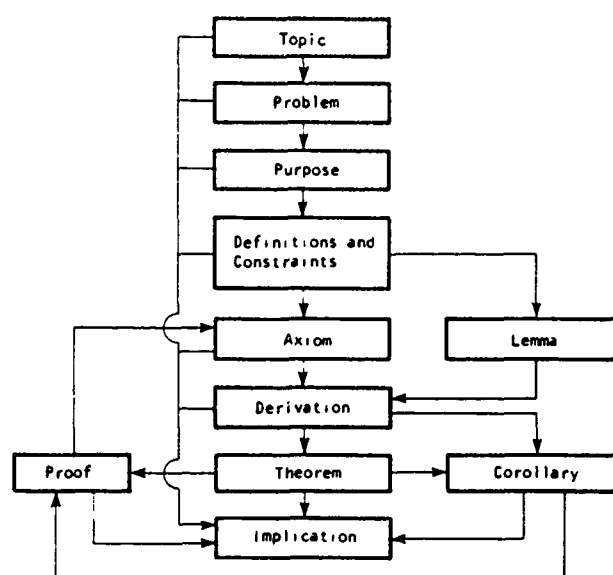


Fig.2 Elements of deductive text

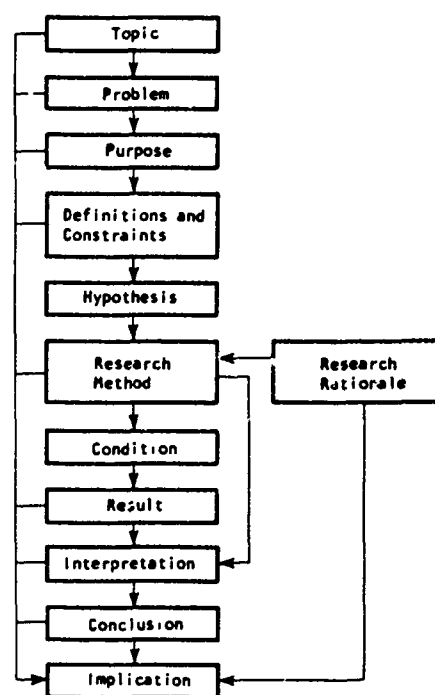


Fig.3 Elements of text describing empirical research

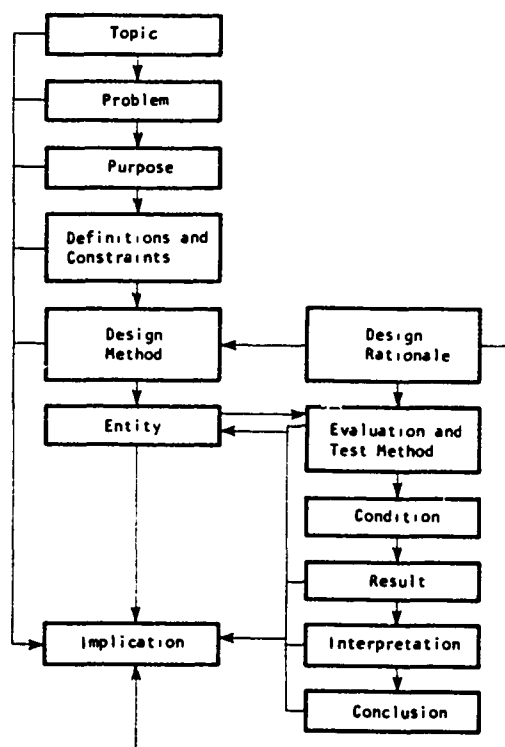


Fig.4 Elements of text describing development and test

DISCUSSION

D. Bosman: People engaged in different disciplines attach different meanings to the same descriptor. To what extent could this be incorporated in the concept of "user profiles" introduced in the paper read by M. Dubon (Paper 3)? How useful would it be?

J.W. Altman: There is a very direct relationship between the meanings attached to descriptors and user profiles. The system described by M. Dubon aims to give a match between descriptors and profile but it could also be used to get a measure of the distance between the profile and the descriptors. A quantitative measure of how close is the information available to that requested.

J.R.C. Licklider: One criticism of user studies is that they have paid too much attention to what users say they need and not enough to what they actually need. Another criticism is that such studies examine existing methods and techniques, whereas users need improved methods and techniques. How does your approach get away from existing methods and techniques? How does it find out what users *need* as distinct from what users *do*?

J.W. Altman: We have established a relationship between text and what users do with it. But if you take the existing text and break it down into units and then manipulate the text units in various ways you develop new configurations not in the original text. It has been found that by eliminating redundancy it is possible to cut the existing text by half without destroying the sense of the original. In this way it is possible to identify principles which could have been applied to improve the text from the users point of view at the time when it was written.

PAPER 7

ON-LINE INFORMATION
STORAGE AND RETRIEVAL

by

Professor N.S. Pywes

The Moore School of Electrical Engineering,
Pennsylvania University, USA

SUMMARY

The components of an automatic storage and retrieval system are briefly reviewed.

The storage process consists of transcription of bibliographic detail, and sometimes abstracts, into machine readable form, generation of a thesaurus, and automatic indexing of documents using this thesaurus. Finally, an automatic process may be applied which generates a library classification system for the collection.

The interactive man-computer retrieval process, which follows the storage process, offers the best potential for improving retrieval effectiveness. The user may search thesauri, classification schedules, catalogues or the documents, in a manner very similar to that employed traditionally in libraries but with far less effort and at much greater speed.

ON-LINE INFORMATION STORAGE AND RETRIEVAL

N.S. Prywes

1. DISCUSSION OF THE PROBLEM

This paper deals with the interrelated issues of pre-processing of documents and the effectiveness of retrieval of documents. Justifiably, the storage and retrieval problem is currently of great concern. The effectiveness in retrieving documents is highly dependent on the amount of labour and process invested in the storage of the documents. Namely, the retrieval is greatly facilitated by storage processing products such as catalogues or storage allocation schemes. These are used in retrieval in referencing catalogues or in following a convenient placing scheme while browsing through shelves of documents. In effect, the problems of storage and retrieval are a single problem. This paper reviews briefly the components of a total storage and retrieval system while referencing relevant developments.

The storage process described includes all the functions which take place in libraries and information centres from acquisition to the placing of the documents in the repository. This process, which includes indexing, cataloging and vocabulary maintenance, demands a great deal of time and expertise. In any one of the large libraries or information centres, there are thousands of monographs and serials that are waiting to be catalogued and indexed. These often lay unused because of the dearth of competent cataloguers and indexers, especially those expert in particular subjects and languages. The increased amount of material which is being circulated soon may require substantial increase in staff. Staff with this competence is extremely scarce; low salaries discourage young people from library work. For these reasons the storage process tends to constitute a serious bottleneck.

On the retrieval side, evaluation tests indicate that libraries and information centres operate at a low, almost unacceptable retrieval effectiveness. The library user requiring specific information is overwhelmed with information, much of which is irrelevant.

The mechanizing of procedures in an information centre or a library does not need any more justification than the notion of mechanizing any other industrial, commercial, or service function. The premise of this paper is that automatic storage processing and on-line retrieval are competitive in effectiveness with manual procedures. The automatic procedures are not especially complex and they can be readily applied.

The automated storage processing discussed here includes the following steps. Citations and sometimes abstracts of incoming documents are first transcribed into machine readable form. Natural language processing of title and abstract results first in a concordance of stem words. The concordance may also provide information about the frequency of stem words. In a semi-automatic process, words may be omitted, added, or various relationships established between words to form an open-ended thesaurus. Then, based on this thesaurus, the incoming documents are automatically indexed. Finally, an automatic process may be applied which generates a library classification system for the collection. Such a classification then represents a scheme for placing documents on shelves, in microforms, or in the computer, as appropriate.

The interactive man-computer retrieval process, which follows the storage process, offers the best potential for improving retrieval effectiveness to the point where infor-

mation storage and retrieval systems become really useful. This interactive process has a number of aspects. An individual can communicate with a central computer through a remote terminal "on-line", i.e., where the terminal is continuously monitored by a central computer. The computer deletes, changes and analyzes the queries and retrieves information in "real-time", compatible with the normal working speed of a human. To fulfill these functions, the computer must have a storage capacity of billions of characters with fractional second access to any information.

In the interactive retrieval process the user may search thesauri, classification schedules, catalogues or the documents in a manner very similar to that employed traditionally in libraries; however, with far less effort and much greater speed. He may, for instance, reference documents by title, author, publisher, citation, subject, or browse through citations or abstracts of documents on a common subject, placed together in the memory of the computer.

Methods and procedures like those described in this paper, such as content analysis, concordance and thesaurus preparation and indexing, which require merely clerical procedures, have been proposed for centuries.⁽¹⁾ They have been opposed by those who believe that manual processing of the document has a "quality" superior to algorithmic processing based on selection of words from the abstract or even from the title. The manual approach has a number of ancillary positions that are contested here. For instance, the manual approach also conveys the notion that the subject term vocabulary needs to be controlled, and that only highly competent persons in specific areas should exercise judgment in regard to adding terms; these positions are contradictory to the approach in this paper. The objective of the procedures described here is to do away with much of the vocabulary maintenance work currently prevalent; especially the notes and instructions directed to indexers and cataloguers which would not be required in an automated system.

2. STORAGE

2.1 The Input of Documents and Content of the Repositories

The repository includes a collection of document representations. Each document is an integral entity in the collection. It may be broken down as shown in Fig.1. The examination and analysis of documents in the storage or retrieval processes is usually conducted in the order from top down of the information shown in Fig.1. Generally, as one proceeds downward in Fig.1, greater depth and, frequently, greater volume of information are provided; however, less frequent access is required to the more voluminous parts.

The upper four boxes in Fig.1 are said to contain *association terms*. These are words or terms such as title, author, subject, etc., which identify single or entire classes of documents. The association terms may convey information about various relationships among respective documents, such as having a subject heading or citation in common, or sequence of events indicated by dates of publications, etc.

The language analysis in the storage process could be based on: (a) entire text, (b) association terms and abstracts, or (c) association terms only. The cost of transcription into machine readable form decreases greatly as the amount transcribed is reduced; however, this also reduces retrieval effectiveness. There is an indication, however, that effectiveness of retrieval based on the subject of the document increases considerably (20-25%) when the transcription of the abstract is added to the transcription of the association terms.⁽²⁾ Language analysis of full text does not seem to improve the effectiveness of retrieval sufficiently to warrant the considerably greater cost of transcription. Also, the content analysis of text requires more complex procedures, including syntactic or semantic analysis.

The document collection is only one part of the information in the repository as shown in Fig.2. The other parts contain directories of the association terms, and stratification of these directories. The directories may be considered to be information about information.

The directories may be generated *a posteriori* from the documents themselves. Namely, the association terms shown in Fig.1 may be extracted automatically from each document as it enters the collection. In this way the concordance of the terms for all the documents may be derived automatically. The aggregate of the various types of association terms then constitute an all-inclusive directory or concordance of the association terms. Further processing then establishes the higher level directories which contain assignment of terms to categories and a variety of relationships among the terms.

The generally prevalent approach to indexing and vocabulary maintenance is that of applying human judgement *a priori*. An example of this approach is the establishment of the Dewey Decimal Classification which has divided the library collection into progressively more specific classes. Using this system, professional indexers in libraries assign subject headings (stated in terms of class numbers) from a controlled schedule to the documents as they enter the libraries. In time, such a classification system must be expanded and revised by the library community to recognize new areas not included in previous schedules. Re-examination and reclassifying of documents already in the collection is then necessary to assign the new subject headings to them.

Figure 2 illustrates the *a priori* and *a posteriori* approaches to generating the directories as opposites. (A variety of mixes of these two approaches is possible.)

Retrieval effectiveness tests indicate that *a posteriori* indexing performs as well as *a priori* indexing; and that the lack of term control in a *a posteriori* indexing does not cause deterioration in performance.^(2,3) This will be further discussed below in connection with the evaluation of retrieval effectiveness.

2.2 Language Processing

The simplest language processing procedure is to analyze a text to recognize and generate stems of words encountered in the input material. This involves recognizing the suffixes of words. A suffix editing procedure for English is described by Stone, et al.⁽⁴⁾ A similar procedure for French has been described by Gardin and his associates.⁽⁵⁾ Similar procedures have been developed by numerous other investigators.^(6,7) More sophisticated procedures including matching stem words against a thesaurus and syntactic or semantic analysis of text may be employed in the automatic indexing and classification as discussed below.

Natural language processing and machine translation research are relevant since many of the algorithms developed there are directly applicable to automatic indexing. However, the systems employing the more complex procedures are highly experimental and in many cases the research has not advanced beyond the theoretical considerations.

2.3 Concordance and Thesaurus Generation

Although completely automatic thesaurus generation procedures have been under development for some time, considerable experience has been accumulated with a semi-automatic approach.⁽⁸⁾ Computer aids are provided, but human intellect is applied to the discrimination and grouping of words. The first step in this process is to use computer aids which accept the transcribed portions of the documents as an input and generate a concordance of stem words. This concordance includes title or abstract words in addition to the other association terms in Fig.1. The computer aids also provide frequencies of occurrence for the words in the concordance.

The first step in deriving a thesaurus may be the elimination of the very high and very low frequency words.⁽⁸⁾ Another step would be the indicating of "broader", "narrower" or "related" relationships between words. Especially important also is the recognition of synonyms. It is necessary to establish such relationships as the documents have been written by many people at different times who use a variety of words to designate similar meanings. Categories may be constituted which contain various instances of word usage;

each such word may be given in context. Another approach is to prepare a separate thesaurus for specific subject areas, where appropriate relationships between words are established in the context of the subject areas.

The thesaurus generation process is similar to the vocabulary maintenance functions in conventional libraries. However, the on-line automated aids may provide suggestions with regard to words and categories which deserve the attention of the individual engaged in establishing relationships among words. For instance, frequencies of terms used in retrieval queries and index terms of relevant documents, which have been retrieved in response to these queries, may serve as a guide regarding association and relationships among terms. Various statistics about frequencies of co-occurrence of terms may be used to combine terms into phrases which will be used in their entirety as a single term in the indexing process. Finally, the automatic generation of a classification, described later, may provide further information about grouping and sub-groupings of terms and respective documents to form progressively more generic subject areas.

2.4 Automatic Indexing

Various automatic indexing approaches and systems have been described by Stevens.⁽⁹⁾ The objective here is to review briefly the simplest procedures which have proved effective. In the most simple procedures, stem words derived from titles or abstracts are considered to be the index terms of the respective documents without reference to the thesaurus at all. This simple process has proved effective for retrieval in situations where a user is satisfied with retrieval of any one or few relevant documents. This method has also proved especially effective in an interactive mode of search where the user may guide the computer in search for relevant material.

Automatic indexing may, however, utilize far more sophisticated approaches. A perusal of the thesaurus for stem words derived from titles or abstracts may result in important indexing decisions. It would eliminate undesired terms, or assign documents to classes or categories. Still, a more complex process may assign term phrases based on words co-occurrences or based on syntactic analysis.

2.5 Automatic Generation of a Classification System and Assignment of Location For Documents

The automatic generation of a classification system in fact groups citations of documents in cells in the memory of the computer, very much as the documents on a common subject are grouped on respective library shelves. The retrieval process then consists of a search of several shelf areas in a large library to find the documents relating to a subject on which information is demanded. A classification system, automatic or conventional, has a dual purpose. It is a methodology for placing like documents together but it is also a retrieval methodology by which one may be guided to the group of "like" documents which deal with the area of his interest. Like conventional classifications, an automatic classification system may be used to put documents away, but *only after* the classification system itself is derived from the documents. Namely, it does not precede the documents, but follows them. The automatic classification process is a follow-up on the automatic subject indexing. It attempts to put together in a cell documents which have most index terms in common.

The scope of this paper does not permit a description of the process for automatically creating a classification system. Various methodologies have been used for this process. These consist of employing statistical techniques,^(10, 11, 12) computing "distances" between documents,^(12, 13) and employing co-occurrence of index terms.^(14, 15) The latter approach is simplest in terms of the complexity of the process and amount of processing required. A collection composed of 4,000 documents with a vocabulary of 6,000 index terms has been processed to date.⁽¹⁵⁾ Experiments are continuing at the University of Pennsylvania with collections of tens of thousands of documents.

It is important to note here that automatic classification may be used not only to complement a coordinate-indexing retrieval scheme, but it also constitutes an alternative to coordinate-indexing. If used in a coordinate indexing system, automatic classification methodology provides a storage arrangement and a directory which greatly speeds up the search and retrieval.⁽¹⁶⁾ As an alternative to coordinate indexing, automatic classification and the arrangement of documents in cells allows the user to direct the computer in its search toward the area of interest. This is further described below in connection with interactive retrieval techniques.

3. RETRIEVAL

3.1 Retrieval by Association Terms

A basic property of an on-line retrieval system is a man-computer language which includes in its vocabulary all the association terms. (See Fig.1) A simple search may be initiated by the user communicating to the system a description of desired information. To "describe" a single or a class of documents, it is necessary to supply in a query the association terms as well as the relationships among the terms. The procedure consists of specifying the association terms of the desired documents and the requisite logical or arithmetic relationships among the terms or among other information elements within the document. It is important that a user at the terminal should be capable of expressing a query in terms most convenient for him. For that reason, ample choice must be given to him to search by various types of association terms, such as, author, publication, title words, accession numbers, references, etc. In addition, he should be able to reference the various directories, such as the thesaurus or the automatic classification, to aid him in selection of terms. Similarly, he should be able to specify, for instance, a generic term to include all the narrower terms which correspond to it. Finally, he should be able to examine the citations which are being retrieved by the system and respond by indicating their relevance to his subject of interest. In these interactions with the computer, the display formats of the computer responses are important to the facility with which a system may be used. These formats are arranged to minimize the user's labour in selecting terms or documents.

On-line retrieval systems may be divided into two classes. The systems which aid user formulation of queries and retrieve respective documents are referred to here as *key word systems*. The second type of systems provides automatic reformulation of the query based on indications from the user of satisfaction or dissatisfaction with the retrieved material. In fact, in this manner the user guides and directs the search of the computer system.

3.1.1. Key-Word Retrieval Systems

An outstanding example of the key-word system is the BOLD system at System Development Corporation, developed by Borko.⁽¹⁷⁾ BOLD utilizes on-line displays which assist the user both in acquiring a mastery of the system itself and in performing guided searches. No language analysis technique is used in BOLD and the indexing is entirely manual. The MULTILIST system at the University of Pennsylvania is another example of a key word retrieval capability based on list processing which facilitates split-second retrieval from large document collections. The MULTILIST system includes both manually indexed (artificial intelligence) and automatically indexed (Physics) collections.⁽¹⁸⁾

BOLD and MULTILIST are representative of typical current systems. With these systems retrieval is easier but the basic content of the query is not altered except at the insistence of the user. Namely, while formulation of the query is assisted by the system, there is no attempt at reformulation based on the results of previous searches.

3.1.2 Interactive Query Reformulating Systems

The procedure in retrieval with a reformulating system may be as follows. A user may desire to search the collection to obtain a bibliography on a certain subject. He would then submit a query to the system consisting of word-stem terms. These terms may be found in directories (Fig.2). The system then will use the automatic library classification which has been generated to find the cell(s) which correspond to the largest number of terms in the query. (Alternately, weights may be associated with the terms and cells are selected which have documents indexed with the maximum total weight of the terms.) The user may then consider a number of citations from the respective cell or cells, and he may indicate acceptance or rejection of certain citations as relevant or irrelevant, respectively. The terms corresponding to the accepted or rejected documents will then be examined by the computer and the initial query may be reformulated. It will include additional terms derived from acceptable documents or it will omit some of the initial terms that are in the rejected documents. Based on the newly reformulated query, a search is repeated, new cells are found, and their content is displayed to the user. This process may continue with the input from the user being primarily the approval or disapproval of retrieved material.

This approach has been experimented with in the SMART Project and the results have been evaluated to determine the effectiveness of this powerful strategy.⁽²⁾ Experiments with this approach have also been conducted by Edwards.⁽⁸⁾

3.2 Evaluation of Retrieval Effectiveness

As has been amply illustrated, there are a great variety of thesaurus generation and automatic indexing strategies as well as of retrieval strategies. It is also quite apparent that the selection of a strategy is very critical to the cost and retrieval effectiveness of the system. An evaluation methodology has been developed to determine retrieval effectiveness of systems.⁽¹⁹⁾ As has been already indicated, increased costs and labour in storage processing may result in improvement of retrieval effectiveness. However, the amount of cost measure as related to the improvement in retrieval effectiveness is very important. Also, for various retrieval applications, different degrees of effectiveness in retrieval are required.

Although tests of retrieval effectiveness have often been seriously challenged on a variety of grounds, two measures of retrieval effectiveness appear to receive wide acceptability.⁽²⁰⁾ One of these measures - the *recall ratio* - is the ratio of the number of relevant documents retrieved to the total number of documents in a collection which are relevant to a search. The other measure, the *precision ratio*, is the ratio of the number of relevant documents retrieved to the total number of documents retrieved in a search.

For a sequence of queries interactively executed in the search, a plot can be made of precision vs. recall.⁽²¹⁾ It is important to point out here that only the conjunction of these two measures is meaningful as an indication of effectiveness of retrieval strategies. The most ideal conditions would be those corresponding to unity recall and unity precision. For instance, perfect recall can always be achieved by retrieving an entire collection; the precision, however, would then be extremely low. On the other hand, if the number of retrieved documents is very small, the precision might be unity, but the recall would be very low. This illustrates that the combination of recall and precision must be considered in the evaluation. A strategy is considered to be more effective if its plot of precision vs. recall is described by a curve closer to the ideal point of precision = 1 and recall = 1. Examination of literature⁽²¹⁾ indicates that in this respect, joint recall precision retrieval effectiveness improves as well chosen, more sophisticated language processing techniques are applied, or as the retrieval process is carried out on-line, interactively employing greater choice of association terms.

4. CONCLUSION

The cost and staffing that are demanded of a library that desires to offer effective retrieval services are currently very large. Many smaller libraries try to use cataloguing and indexing material generated in large information centres but even utilization of such resources requires considerable staff and cost. These smaller libraries may be the real beneficiaries from a total on-line storage and retrieval facility as described in this paper. The state of the art indicates that such a system is feasible and economical to develop at this time.

ACKNOWLEDGEMENT

The work discussed herein was supported by Contract NOnr 551 (40) from the Information Systems Branch, Office of Naval Research and Rome Air Development Centre.

REFERENCES

1. Sayers, W.C.B., *A Manual of Classification for Librarians and Bibliographers*. 2nd Ed., Grafton & Co., 1944
2. Salton, G., Scientific Report No. ISR-11 and No. ISR-12, Information Storage and Retrieval, Cornell Univ. Dept. of Computer Sci., June 1966 and June 1967.
3. Cleverdon, C.W., et al. *Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Part 1: Text*. Cranfield, U.K., ASLIB Cranfield Research Project, 1966.
also
Cleverdon, C.W., *Report on Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Cranfield, U.K., ASLIB Cranfield Research Project, Oct., 1962.
4. Stone, P.S., et al. *The General Enquirer: A Computer Approach to Content Analysis*. Cambridge, Massachusetts, The MIT Press, 1966.
5. Gardin, J.C., *Syntol, Vol. 2*. New Jersey, Rutgers State University, 1965.
6. Salton, G., *Content Analysis*. Paper given at Symposium on Content Analysis, Pennsylvania Univ., Nov., 1967.
7. Sager, N., *A Syntactic Analyzer for Natural Language*. Reports on the String Analysis Program, Pennsylvania Univ., Linguistics Dept., March, 1966, pp. 1-41.
8. Edwards, J.S., *Adaptive Man-Machine Interaction in Information Retrieval*. Ph.D. Thesis, Pennsylvania Univ. Moore Sch. of Electr. Engng., Dec. 1967.
9. Stevens, M.E., *Automatic Indexing: A State of the Art Report*. Monograph 91, National Bureau of Standards, U.S.A., 1965.
10. Borko, H., *Research in Automatic Generation of Classification Systems*. Proc. Spring Joint Computer Conf. 1964, pp. 529-535.
11. Williams, J.H. Jr. *A Discriminate Method for Automatically Classifying Documents*. Proc. Fall Joint Computer Conf. 1963.

12. Baker, F.B., *Information Retrieval Based Upon Latent Class Analysis.*
J. Ass. computing Mach., Vol.9 No.4, Oct., 1962, pp.512-521.
13. Needham, R.M., *Automatic Classification in Linguistics.* California, Rand Corp. Rep., Dec., 1966 (AD 644 961)
14. Prywes, N.S., *Browsing in an Automated Library Through Remote Consoles*
in Sass, M.A., Wilkinson, W.D., *Computer Augmentation of Human Researching* (Proc. of Seminar, June 1964) U.S.A., Spartan Books, 1965, pp.105-130.
15. Lefkowitz, D., *Experiments in Automatic Classification.* Computer Command
Angell, T., and Control Co., U.S.A. Report 85-104-6, Dec., 1966.
16. Prywes, N.S., *Structure and Organization for Very Large Data Bases.*
presented at Symposium on Critical Factors in Data Management,
1968: Problems and Solutions, California Univ., U.S.A.,
20-22 March, 1968.
17. Borko, H., *Design of Information Systems and Services.* in: American
Doc. Inst. *Annual Review of Information Science and Technology*, Vol.2. New York, John Wiley & Sons, 1967.
18. - A collection of Physics Articles prepared by a project at
the Massachusetts Institute of Technology under the direction
of M. Kessler. The experiments conducted with this collection
are a subject of a Master's thesis, *Automatic Introduction
of Information Into a Remote Access System: A Physics
Library Catalog*, by P. Gabrini at the Moore School of
Electrical Engineering, (Pennsylvania University, Report
67-09) December, 1966.
19. Bourne, C.P., *Evaluation of Indexing Systems.* in: Cuadra, C.A. (Ed.)
Annual Review of Information Science and Technology, New
York, John Wiley & Sons, 1966.
20. Swanson, D.R., *The Evidence Underlying The Cranfield Results.* Lib. Quart.,
Vol.35, 1965, pp.1-20

also

On Indexing Depth and Retrieval Effectiveness. in: Spiegel,
J., Walker, D. (Eds.) *Proceedings of Second Congress on
Information System Sciences.* U.S.A., Spartan and McMillan,
1965.
21. Salton, G., *Computer Evaluation of Indexing and Text Processing.* J.
Lesk, M.E., Ass. comput. Mach., Vol.15., No.1, Jan.1968, pp.8-36.

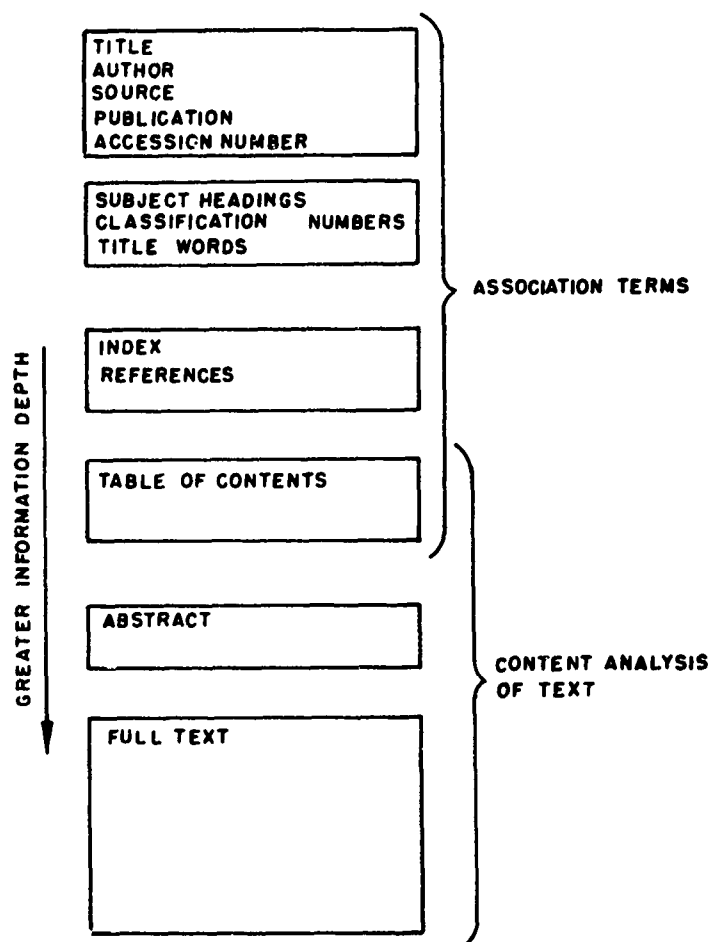


Fig.1 Hierarchical breakdown of a document

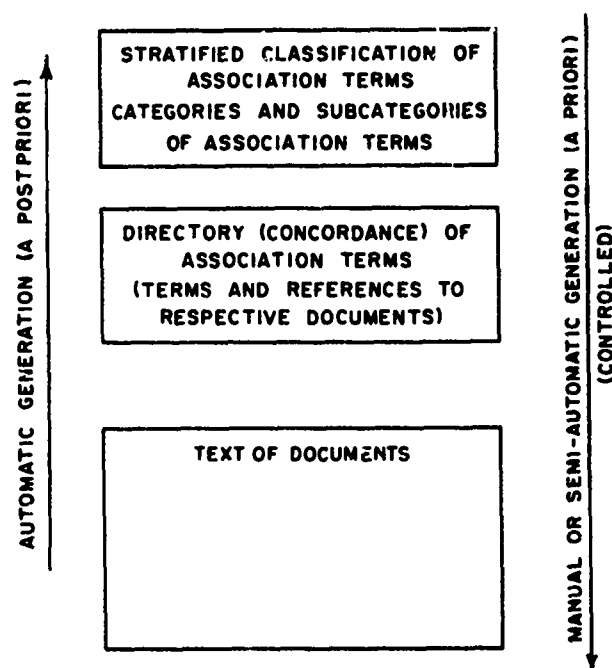


Fig.2 Hierarchical breakdown of information in a repository

DISCUSSION

J.R.Weiner: What precision have you obtained in your retrieval, and what do you hope to attain?

N.S.Prywes: My work is with a large collection of documents and satisfactory tests for precision of retrieval from large collections are still required.

Professor Salton, working with a smaller collection (Reference 2) has obtained precision of greater than 60 per cent.

H.F.Vessey: I am disturbed at the statement that terms occurring infrequently are eliminated in thesaurus compilation. Terms such as project titles, names etc., might only occur once or twice but be very powerful terms for retrieval purposes.

N.S.Prywes: The list of low frequency words is comparatively short, only a few hundred words. This list could be weeded manually to retain such terms as might be useful in a search.

A.H.Holloway: You have mentioned a precision of about 20 per cent and that Professor Salton expects to achieve 80 per cent but have not mentioned the recall. It is not difficult to achieve a precision of 100 per cent with a very low recall. Can you say what combination of these criteria you hope to achieve?

N.S.Prywes: We are investigating whether our users want good recall or good precision as alternatives. For teaching and research it is often acceptable to have high precision with fairly low recall.

R.Bree: Could you please say:

- (1) The number of documents used in the trials of the system.
- (2) From what part of the text is the descriptor material extracted.
- (3) What is the computer economy of this method of mechanical text analysis.

N. S. Prywes:

- (1) A collection of 6,000 documents obtained from the Department of the Air Force.
- (2) About 10,000 words have been extracted from the titles only of the document.

We are proposing to test the system on a larger collection of documents in Nuclear Science Abstracts, using titles, Universal Decimal Classification headings and Defence Documentation Centre Descriptors.

- (3) Systems must be worked on a serial process such as magnetic tape.

The classification has been organised into a tree at three levels. It takes three passes of tapes for the entire collection to obtain material for each level. Each pass takes about one hour. Updating is carried out monthly and takes about ten hours. Daily updating was tried but this did not produce sufficient new entries.

PAPER 8

NON-NUMERICAL MATHEMATICS
AND DATA PROCESSING

by

F. Krückeberg

Bonn University, Germany

SUMMARY

Several problems involving non-numerical mathematics are listed. In the field of non-numerical data processing, the following topics are discussed briefly:- Group theory; Games theory; Translation; Graph theory; Pattern recognition and enhancement.

NON-NUMERICAL MATHEMATICS AND DATA PROCESSING

F. Krückeberg

1. NON-NUMERICAL MATHEMATICS

Many problems of modern mathematics are non-numerical in structure. It is a simple matter to calculate the path of a rocket numerically, based on the theory of differential equations. This subject properly belongs to Analysis. On the contrary, the topological structure of cyclic satellite orbits is non-numerical. It is possible to classify types of orbits topologically^{1, 2}. (See Fig.1).

There are mathematical areas which contain no numerical components as, for example, graph and network theory. With the help of these theories, complicated problems of strategy can be analysed. In the theory of games, graphs allow an intuitive grasp of the problem to be readily achieved. There is a close correspondence between graph theory and logic. Graphs can also be described in terms of Boolean matrices. With this, a link to algebra emerges. A special topic in algebra is Group Theory which can be used for the investigation of graphs. The importance of Group Theory is, however, much greater than this and more general. For example, one can with the help of groups, describe the symmetric properties of elementary particles.

Many problems in geometry are non-numerical in nature. Hilbert's research on the foundations of geometry are especially worthy of note here. A further important subject in mathematics is logic. This topic is, at present, being very actively pursued. It is possible to prove that large classes of problems can be solved without direct consideration of the individual problems through the utilisation of very broad logical generalities. An extension of this idea leads to the new topic of model theory. Modern mathematical theory is becoming ever more generalised and distant from the classical world of numerical analysis.

2. NON-NUMERICAL DATA PROCESSING

The field of non-numerical data processing is so large, if one regards it with full generality, that no list of subjects can be exhaustive. One can merely cite several important new efforts without prejudice to the large number of others not mentioned.

2.1 Group Theory

It is possible to store finite groups in the core store and all group operations can be described in subroutines. In this way it is possible to manipulate groups in computers. For example, all sub-groups can be determined automatically and even more complicated problems of group theory can be solved very conveniently^{3, 4}. In this domain, it is certain that very interesting new results will be obtained. (See Fig.2).

2.2 Games Theory

It is well known that it is possible to program a computer to play a complex game like chess. Since the theory of games is of the greatest importance for scientific management and logistics, game playing in the computer becomes a very serious occupation indeed. This application of the computer will be, in future, one of the most important. A special possibility is the construction of time-tables by computers⁵.

2.3 Translation

Machine translation is of considerable importance due to the continued growth of international scientific exchange.

In the area of language data processing the German "Forschungsgruppe LIMAS" (Forschungsgruppe Linguistik und maschinelle Sprachübersetzung) has shown how flow diagrams can be used for the purpose of machine translation. Here language is viewed as a process, i.e. a system, in which information is converted into speech signs. This process operates on three main levels.

The first level is called the "nomo-sphere". It is composed of "Inhalt"-factors ((semantic regulating factors)(semantische Steuerungsfactoren)) which alternatively integrate with and modify one another.

The second level is a parallel system which portrays the "morpho-sphere" of the language. The above portrayal is brought about by "formale" rather than "Inhalt"-factors. Linear cohesions or single factors or factor groups seldom occur between these two levels.

Cohesive relationship between levels 1 and 2 are built by a system of combinations and ramifications. This level is called the "nomo-morpho interaction bridge" (Wechselwirkwerk).

All cohesions at all three levels - the morpho-sphere, the nomo-sphere and the morpho-nomo interaction bridge - are linguistically determined and selected.

A structural picture of such a system of relationship at the morphospheric level is illustrated in the accompanying flow diagram. It is an example of the system of relations existing between all three levels. The range between these two levels morpho-sphere and nomo-sphere is the nomo-morpho interaction bridge. This structural picture appears also in all subprograms.

The function of such programs is reversible and the nomo-factors and the morpho-factors are information carriers.

The goal of this system is to portray a formal image of the communications process called "language, i.e. with the same grammar and flow system synthesis and analysis speech and understanding can be carried on.

Dr. Hoppe calls his system "Kommunikative Grammatik". It is characterised by the following principles:

- (a) the reversibility of functions
- (b) the information-retention of the regulating factors
- (c) the functioning of the factors
- (d) the integrating of the factors
- (e) the binary structure of the functions
- (f) the nomo-morpho interaction bridge
- (g) determination
- (h) selectivity
- (i) the operation of the process in time.

Such a non-numerical treatment of data belongs to the area of data processing, not to non-numerical mathematics. This treatment contains a working theory which allows, without the help of logistical functions, the generation and transformation of the process, the

explication of factors which are not morphologically represented and the verbalisation of these factors, that is, their expression in grammatical forms.

In this way a system for the direct coordination of signifier and signified is replaced by a complex system of functions of numerous semantic and formal regulating factors.

Only when language is treated in this way, as a regulated process and as a system of functions, is the way paved for high-quality machine translation.

By means of the above mentioned factor characteristics (information carrier, binary functionality, reversibility, reciprocal integration, selectivity, determination) a great number of the so far unconquered problems can be solved in the process of translating, problems among which the ambiguities, the indefiniteness and the implied information are the most important.

Translation connects the factor-process-system of two languages by way of a factor formula which contains the regulating principle for each of the sentences to be translated, as it is machine-analysed in the input language and as required for synthesis of the target language in its regulating process. The translation is in this case equivalent in meaning (to the original), however it need not always be comparable in form, that is, in its syntax. The LIMAS-system has already been thoroughly explained in a number of publications^{6, 7, 8, 9, 10, 11}. (See Fig.3).

A Russian-German translation project is being carried out in collaboration with the Deutsche Forschungsgemeinschaft (German Research Association) and the University of Saarbrücken (See "Systran System" - P. Toma¹²).

2.4 Graph Theory

Very complicated graphs can be stored in the computer, and the structure of the graph can then be investigated. It is possible, for example, to determine the shortest connection between two nodes of the graph. Further, cyclic sub-graphs can be discovered. Such questions are of the greatest practical interest. Techniques, including signal-flow graphs and k-trees allow one to obtain a clear intuitive picture of the functioning of a linear electrical network, after which analysis is much easier. Knowledge of the graph in topological network analysis eliminates many time consuming mesh and node calculations. In the chemical industry, the flow of material can be described by these methods. This is of considerable importance for the solution of management problems in such large factories. A large German chemical factory is, at present, actively using this technique in its daily operation¹³.

2.5 Pattern Recognition and Enhancement

The recognition of shapes is a very difficult and interesting problem whose solution has many applications. The problem can be divided into two parts. One is the decision as to which class out of a large number of possibilities a given well defined pattern belongs (character recognition, for example). The other is concerned with improving the definition in patterns which are greatly disturbed by extraneous influences and, given a limited number of classes, deciding into which such patterns most probably fall. For example, in the latter category, in collaboration with the Rheinisches Landesmuseum, Labor für Feldarchäologie, a project on the enhancement of buried archaeological monuments seen in the results of surface geophysical measurement is in progress. Although the method requires much numerical manipulation of the data, the end result must be presented in a form which enhances the ability of the human eye to distinguish faint shapes in a noisy field¹⁴. (See Fig.4).

Form, pattern structure, logic trees, graphs, language, algebraic manipulation, all of these are but a few of the non-numerical problems which are yielding to the attack of non-numerical mathematics and data processing, giving new results in areas where hitherto insurmountable difficulty prevailed.

REFERENCES

1. Arenstorf, R.F., *A New Method of Permutation Theory and its Application to the Satellite Problem of Celestial Mechanics.* J. reine angew. Math., Vol.221, 1966, pp.113-145.
2. Arenstorf, R.F., *New Periodic Solutions of the Plane Three Body Problem.* Presented at Int. Symp. on Differential Equations and Dynamical Systems, Mayaguez, Puerto Rico, Dec.1965.
3. Gerhards, L.,
Lindenberg, W., *Ein Verfahren zur Berechnung des vollständigen Untergruppenverbandes endlicher Gruppen auf Dualmaschinen.* Num. Math. Vol.7, 1965, pp.1-10.
4. Lindenberg, W.,
Gerhards, L., *Combinatorial Construction by Computer of the Set of All Subgroups of a Finite Group by Composition of Partial Sets of its Subgroups.* Proc. of Conf. on Computational Problems in Abstract Algebra, Oxford, England, 1967.
5. Genrich, H.J., *Die automatische Aufstellung von Schulstundenplänen auf relationentheoretischer Grundlage.* Schr. Rhein.-West. Inst. Instrum. Math. Univ. Bonn, Ser. A, Vol.11, 1966.
6. Hoppe, A., *L'explication automatique du facteur sémantique "appartenance" apparaissant dans le "Zuwendsatz" allemand, et la traduction automatique de ce facteur.* Tagber. Conf. Int. Traitement Autom. Langues, Grenoble, 1967.
7. Schweisthal, K.G., *Demonstration of a Ring-Model for German/English - English/German, Including Analysis-Synthesis, Transformation and Generation.* Tagber. Conf. Int. Traitement Autom. Langues, Grenoble, 1967.
8. Engelen, G., *Programming of Reversible Systems in Computational Linguistics.* Tagber. Conf. Int. Traitement Autom. Langues, Grenoble, 1967.
9. Schnelle, H., *Über den Stand der Forschung zur automatischen Sprachbearbeitung im deutschen Sprachraum.* Beitr. Linguistik Informationsverarb. Heft 2, 1963, pp.48 ff.
10. Zint, I., *Über den gegenwärtigen Stand der automatischen Sprachbearbeitung.* Beitr. Linguistik Informationsverarb. Heft 12, 1967, pp.36 ff.
11. Mey, J., *On the Present State of Automated Language Processing.* Computg. Rev. 1968, p.316.
12. Toma, P., *The Systran System.* Germany, Saarbrücken Univ.
13. Wenke, K., *Mathematische Analyse von Abhängigkeiten.* in Festschrift Carl Wurster, BASF, Ludwigshafen/Rh., 1960, pp.421-433.
14. Scollar, I.,
Krückeberg, F., *Computer Treatment of Measurements From Archaeological Sites.* Archaeometry, Vol.9, 1966, pp.61-71.

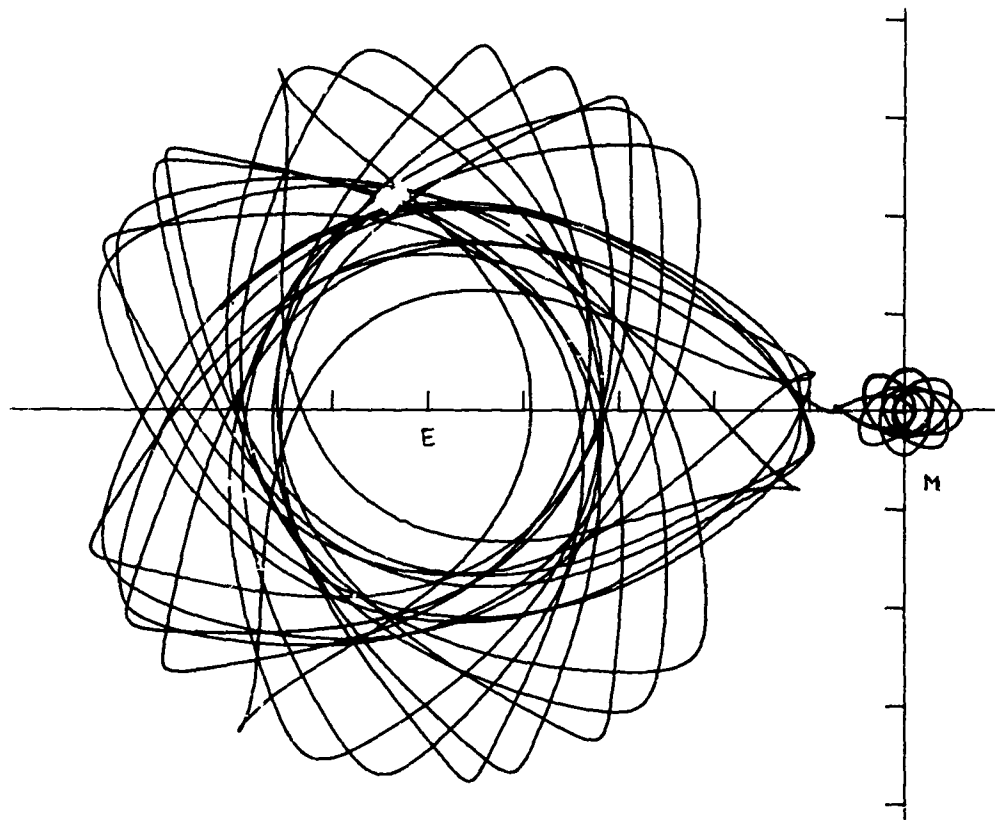


Figure 1. Orbits of a satellite between earth and moon

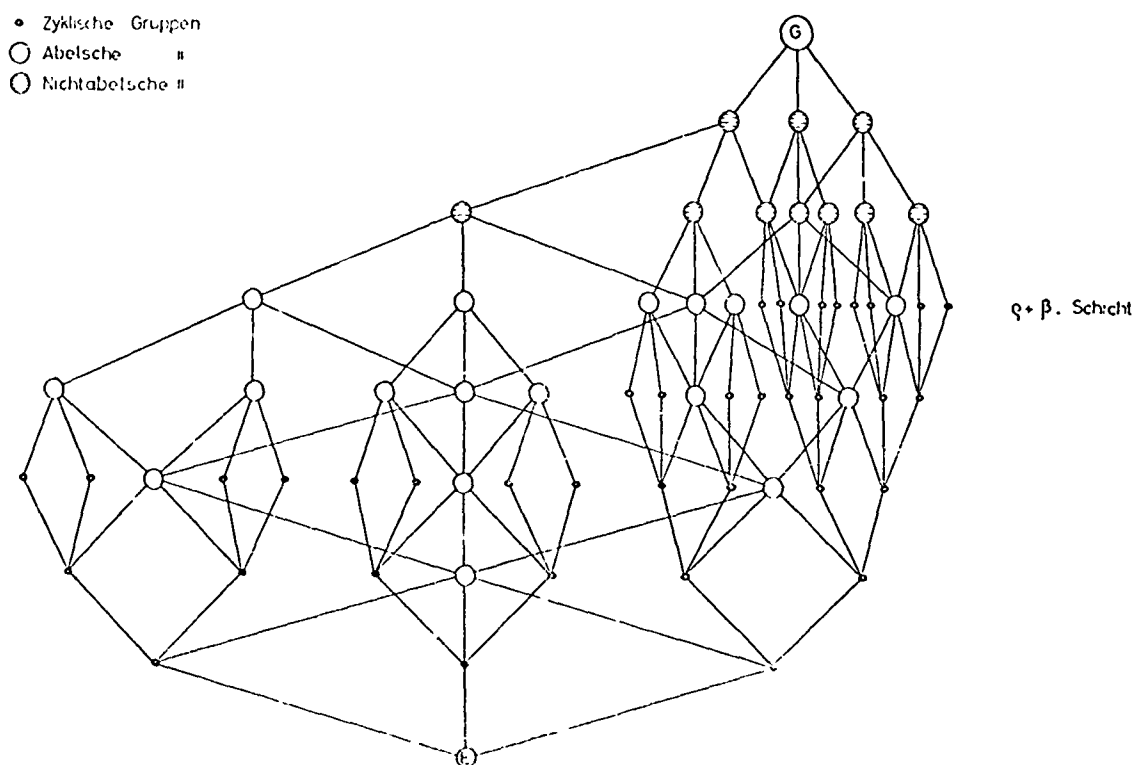


Figure 2. Lattice of all subgroups of a group of order 2^8

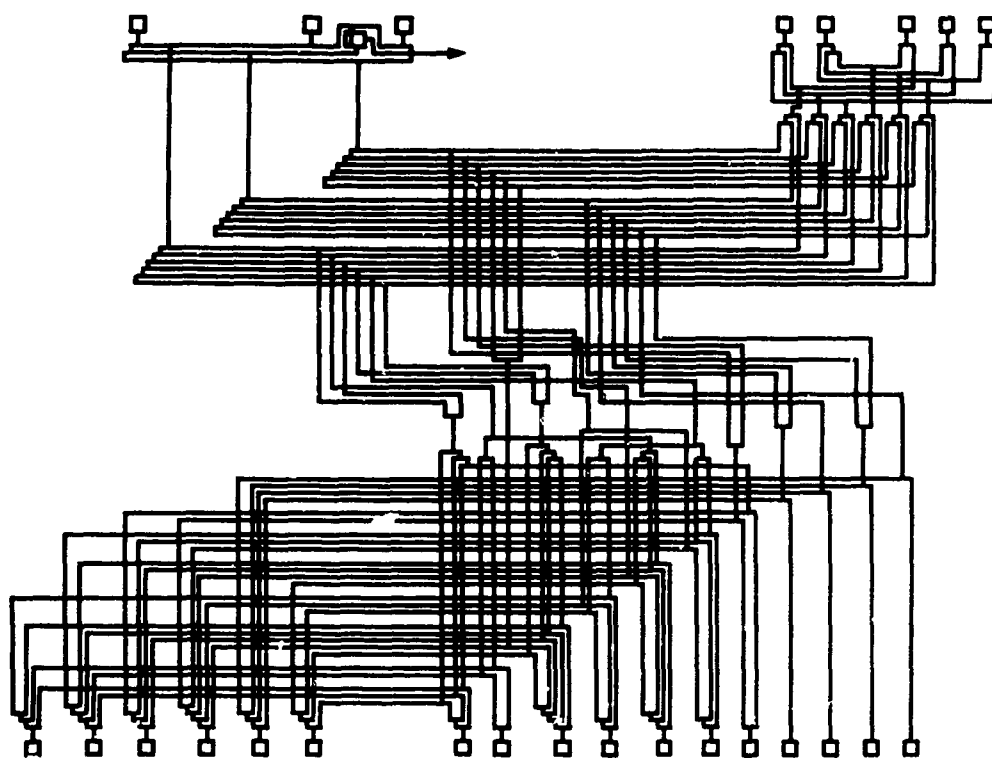


Figure 3. LIMAS flow diagram

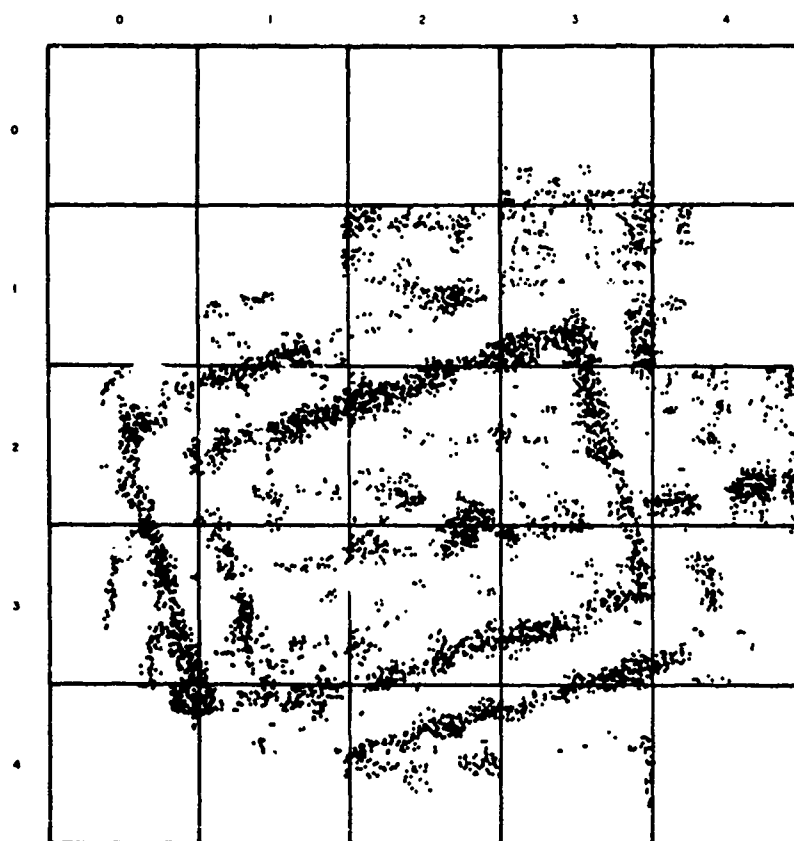


Figure 4. Walls of an old Roman settlement

DISCUSSION

Lustig: I would like to raise three points:-

- (a) Language translation does not seem to be a very good example of non-numerical mathematics.
- (b) Non-numerical mathematics seems to have a very limited use in the documentation field.
- (c) There seem to be many areas of mathematics which are amenable to computer solution but this does not seem to be a common practice.

F.Krückeberg: It is probably true that the use of non-numerical mathematics in documentation studies is limited at present but more extensive use may be made of the technique in the future.

Regarding the solution of mathematical problems, there is plenty of scope for the application of computers; at Bonn University many mathematical problems are solved in this way already.

N.S.Prywes: Can you comment on the use of graph theory methods for simplifying a diagram of a thesaurus.

F.Krückeberg: Established theories and methods exist for reducing graphs and electrical networks and these may be applicable to this problem.

PAPER 9

MANUAL SYSTEMS -
TDCK CIRCULAR THESAURUS SYSTEM

by

J.A. Schüller

TDCK, Netherlands

SUMMARY

The organization, functions and systems used at TDCK are described. TDCK collects, evaluates and stores information primarily useful for military purposes. The retrieval systems used are Universal Decimal Classification and the TDCK-Compact System. The TDCK Thesaurus has been designed such that related concepts are placed on concentric circles; arrows, fanning out in all directions, are used to display relationships between the descriptors. At the "input" coding will start from the centre of the circle, following an arrow until the wanted descriptor is reached. A total of 376 circle-schemes have been designed so far.

MANUAL SYSTEMS - TDCK CIRCULAR THESAURUS SYSTEM

J.A. Schüller

1. ORGANIZATION OF TDCK (The organization is shown in Fig.1.)

The centre falls directly under the Minister of Defence. An advisory council consisting of four members advises the Minister, at his request or on their own initiative, on TDCK-policy matters.

Three of the members are high-ranking officers of the Navy, the Army and the Air Force and the fourth member is the director of TDCK.

TDCK personnel strength is 63 people.

Following the scheme of Fig.1 we see at the left-hand-side the technical divisions and at the right-hand-side the special-library department and the administration division.

The technical division, subdivided in sections, is manned by scientists and technical engineers, in total 25, nineteen of whom hold an academic master's degree, while the remainder have bachelor degrees or are senior serving officers.

2. FUNCTIONS

The primary functions of TDCK are:

- (a) To collect, evaluate and store new scientific information, from all over the world, which may be useful for military purposes in general.
- (b) To be well informed of highly specialized information-sources in-and outside the country.
- (c) To use available information for giving assistance to those scientists, technical investigators and officers, who are involved in solving problems in research, technology, education, management, military sciences and other fields of military interest.

Before explaining how we try to fulfil our functions, and describing some details of specific activities, I should like to stress that obviously a documentation centre is a model for applied efficiency. Its charge is to use available information, to inhibit costly duplication and to select the most effective and efficient modern methods for achieving these aims.

The most simple and direct definition of documentation which has come my way lately runs thus:

"Documentation embraces the logistics of knowledge"

Whether we achieve this communication or transmission of information by handsorting activities - more or less mechanized - or by a computer, does not affect the functions of a documentation centre.

The first-mentioned function, collecting new scientific and technical information, includes the evaluating factor and both are dependent on the fields of interest, and the degree of specialization within these fields, of the institution for which one is working. It may be clear that TDCK's charge: "for military purposes in general" encompasses a very broad area, and I refer again to Fig.1 where this coverage is indicated.

Collecting scientific and technical data useful for defence purposes is one of the intricate activities of the centre. Firstly one should know what is wanted; secondly what is to be had and where.

After some 15 years TDCK has received reports from more than 6000 different research institutes working for defence and spread over the 15 NATO-nations. These include all types of report-producing institutions; e.g. laboratories of research establishments, of universities and of industries. In certain cases TDCK has succeeded in being put on mailing lists, which is exceptional for a foreign centre. Other institutions send their accession lists from which reports may be requested. In many instances TDCK receives technical reports without cost, but often also in exchange for TDCK publications, while other series are made available to TDCK against reproduction costs. The bulk of the reports acquired by TDCK are received from the United States, but significant contributions come from the UK, Canada, W-Germany and AGARD. TDCK maintains close contacts with the national defence documentation centres in NATO countries; contacts which are encouraged by the AGARD Technical Information Panel membership.

A network has been constructed which actually connects the information-centres in NATO and occasionally in some neutral friendly countries. At the same time direct contacts are maintained with several special institutions in some of these countries (see Function 2(b)) resulting in a regular exchange of reports.

In order to minimize duplication of research, TDCK will buy keys the research literature where possible. It subscribes to "Physics Abstracts", to "Environmental Effects on Material and Equipment Abstracts" to "Excerpta Medica" to "Index Aeronauticus" to "Meteorological and Geostrophysical Abstracts", etc. In total we have over 30 different subscriptions of this kind. These abstracts are considered to be the backbone of our information sources, and only supplementary work is needed, keeping in mind that many of these abstracting services are not up to date - running behind from, say, three months to two years - and moreover do not cover all publications e.g. symposia-papers, patents and unpublished reports.

Interesting new articles and many unpublished documents culled from many sources are selected, abstracted, and published in our monthly literature digests. Each scientific section at TDCK composes its own digest so that we publish 20 different literature digests per month, three of which are issued every two weeks (namely "Electronics", "Aeronautics", and "Economics"). Perhaps I should mention also that in some subject areas we are working in co-operation with other documentation centres, in so-called "Pools" preparing abstracts for common use; our unclassified digests are also circulated to interested parties outside defence.

All reference material published in our literature digests is entered in our indexing and retrieval systems and is available in hardcopy. For defining the contents of the literature and for specialized retrieval search TDCK is using two different systems: namely the UDC (Universal Decimal Classification) and the TDCK-Compact System.

Apart from these systems an index is kept for retrieving reports, papers etc. according to their issuing organizations; an institute, a laboratory, etc.

In some cases the information scientist, or the questioner, is aware of specific long-term research projects which are undertaken by one or more well known laboratories. In such cases he may find useful data at short notice in this index. For example the answer to a question concerning: "theory on 3-4 or 5 bladed supercavitating propeller performance", may

be quickly found under TMB (Taylor Model Basin Wash.DC), since it is known that TMB is working in this field.

Part of this retrieval activity is mechanized by a Lectriever, which is a mechanically driven file selector.

Now I propose to turn to figure 2. Here is shown our Table to Sources of Information, a list which is intended to remind our information officers of all sources *available at TDCK* which should be consulted when, for instance, a selective bibliography has to be compiled.

Along the top are found (vertically printed) the different scientific and technical areas which are of interest to the M.O.D.; i.e. the same broad subject areas as shown in Fig.1.

In the column at the left we distinguish three different groups of information sources:

1. Card Catalogue Systems
2. Books of Reference
3. Abstracts Journals.

If, for example, a request for a bibliography on "Inertial Navigation Systems" has been requested, 18 different sources will have to be consulted according to this schedule all of them packed with information and all of them using their own indexing system. This last remark suggests one of the reasons why TDCK does not feel like using a computer for its own system; some 35 other systems would still have to be scanned in a conventional manner.

Coming to the third function of TDCK we arrive at the main objective for which a defence documentation centre is established, namely to provide information to those who need-to-know. Technical questions which are put to TDCK are handled according to the needs and the wishes of the questioner only, of course, as long as TDCK is able to provide specialists time and material. In principle a question can be answered in four different ways:

- (a) By making available a selection of reports or articles in which the problem has been treated.

If the questioner knows the title of such reports the request is a very easy one to handle and is limited to routine library action; if not, the centre has to find the required information by one of its retrieval systems.

- (b) By making available a bibliography consisting of titles and descriptive abstracts of all available printed information, all well indexed and cross-referenced.

When such a bibliography has to be compiled, all available sources in TDCK have to be consulted (see Fig.2.)

- (c) Through the production of a literature research report or a "state of the art" review, in which the information scientist provides a survey of the latest developments in the relevant subject area. Requests for such studies arrive more and more frequently.

The information centre of today is already manned by a university trained staff of engineers and doctors with linguistic abilities, for selecting the literature, for making descriptive abstracts and for classifying.

It is a very attractive part of the literature analyst's job to make literature searches, and is undoubtedly a highly responsible scientific task which - in my opinion - will never be accomplished by a computer.

Of course these extensive special studies can only be made when enough time is available.

- (d) By offering research workers a chance for interaction with the literature in their specific narrow field of interest. In other words providing facilities which permit effective browsing; in my opinion a very necessary activity.

The number of complex questions which entailed much work exceeded 700 in 1967. Less intricate questions amounted to about 1400, while requests for copies of specific reports or articles totalled more than 50000 during last year.

3. SYSTEMS

Coming to the systems which TDCK uses for indexing and retrieving the literature by subject, I should like to give you an idea of the philosophy of TDCK's so-called Compact System.

Before doing so I have to try to convince you that a documentation and information centre is constantly confronted with very specific difficulties. A well known example is that a visiting scientist research worker frequently does not know what he is actually looking for, and finds himself unable to formulate his information need clearly. The only thing we can do in such a case is to confront him with a scientist from our staff, a colleague who understands his language (his jargon), if not his problem, and who is able to lead the enquirer to a manual system in which he can browse. The big questions in information systems - on the documentation side - are always "will we get out all that we have put in?" and, "did we put in properly what we have?"

For many years TDCK has used two different systems for retrieval and, to anticipate a logical question, I want to stress that there is no system in the world today which will give a 100 per cent output. The application of two systems which differ fundamentally in their nature and philosophy, puts at our command the sum of the possibilities inherent in each of the two systems.

Of course this means too that more work has to be done.

Experience, however, indicates that this "more work" has to be done on the input side and that, in most cases, one will meet with less work and certainly more completeness on the output side. Moreover, when handling two systems for indexing all documents, it is possible to pose identical questions to both systems and to learn why a certain document was not retrieved by one of the two systems. In this way research will pinpoint the weaknesses of both systems.

The results of such research has led in our case, to the design of a new system which should replace the uniterm-system of coordinate indexing, which has shown some serious deficiencies.

This new system is called the TDCK Circular Thesaurus System. It was considered that certain features of several well-known systems are very useful and, when possible, should be incorporated in our new thesaurus conception.

So most of the ideas used in the TDCK-system are not new at all. In fact, what is new is that, if properly used, the visible display of a systematically built thesaurus compels a person to retrieve with the same terms as used at the input activity. This aspect is perhaps the most significant one of the TDCK manual system; the graphical display of the scientific sub-divisions of a discipline introduces a third dimension to the thesaurus. When designing this thesaurus we consciously sought to obtain a combination of:

1. a systematic subject set up;
2. alphabetical arrangement of descriptors;
3. coordinate indexing principle;
4. mutual relations and facets; and finally
5. visible directions display.

We believe that in the new TDCK Circular Thesaurus method such a combination has been achieved.

I shall now try to give a description of what we have called a simple circle-scheme.

Notions which are related to each other, and which can be placed in the familiar family-tree pyramid, are placed on concentric circles; that is to say we have simply made a plan-view of our tree because on a circle we have, in practice, about five times more room than in a pyramid structure. Arrows, fanning out in all directions are used to display relationships between concepts. An example of a circle scheme is shown in Fig.3; in use a specific procedure is practised - this is described in some detail below. Briefly when arrows are followed from the origin of a circle-scheme we see how notions fall into logical subdivisions. A relationship is sought which, depending on needs can be continued on a following concentric circle, and so on.

When descriptors from other circle-schemes have to be used for describing a document properly, such notions can be "borrowed", but should immediately be added (in writing) to the circle-scheme of current interest, but outside the circle. All such "complemented" circle-schemes are formally published.

The individual circle-schemes which have been built up are thus kept limited, and arrows will refer us to other circles when we are entering their domain. It may be observed also in Fig.3 that in all cases *these* arrows point to "borrowed" descriptors which are not framed. The word at the centre of each circle-scheme is usually a descriptor with a high frequency count in the system.

The following rules are in force:

1. The thesaurus consists of descriptors;
2. Each framed descriptor appears only once in the system;
3. At the *input*, coding will start from the centre of a circle, following an arrow until the wanted descriptor has been reached. In one circle, more than one radius may be followed;
4. All descriptors encountered will be noted down.

When a descriptor from another circle has to be used, we can "borrow" such a notion and add it to our circle-scheme, outside the *ultimate* circle, in which case we do not use a frame.

5. New descriptors have to be defined by the subject-specialist concerned. They will be added officially to one of his circle-schemes after which they can be "borrowed" by other circle-schemes;
6. On the *retrieval* side the circle-schemes will always be used. The user (a specialist in the field) will be led automatically to the pertinent descriptors, once the appropriate scheme has been selected.

Only two or three descriptors are necessary for defining the question, in other words: a document coded by 20 descriptors can give an answer to 10 different questions.

7. The thesaurus, as a one-language technical index, can be translated.

The use of homonyms and synonyms is avoided. The word "measurement", for example, will be used in several descriptors, which could read: measurement of time, measurement of distance, ballistic measurement, etc. In all, 376 circle-schemes have been designed to date.

The number of descriptors for the TDCK fields of interest (defence) is about 11,500. It is expected that this number will gradually increase to no more than 12,500. Accessions of essential descriptors are reported to the System-Manager, who publishes, usually every year, up-dated schemes to replace old ones. The fourth edition of the TDCK Circular Thesaurus was

published in 1966, the fifth edition is in print at this moment; it will contain almost 400 separate circle-schemes. It is easy to re-arrange any scheme, if this is desirable or necessary, as for example when the philosophy in a particular scientific field is subject to change. The sequence of the descriptors along one radius, however, is never subject to change.

Summarizing: Hierarchically related thesaurus terms are arranged within a series of concentric circles, with the most generic term at the origin. Arrows radiate outward to specific terms on the first circle and from some of these terms to successively more specific terms on succeeding circles.

Fig.4 is taken from page 100 of the Thesaurus where the division *Operational Research* has been subdivided in 15 so-called "descriptor fields".

If, for example, we receive a document in which the aspects of a tactical air defence O.R. game are discussed it will of course be passed to our O.R. division for abstracting and indexing. Here the proper descriptor field will be chosen (Fig.4). Our specialist will turn to scheme 112 **Operational Game** and on this page the descriptor field shows a plan view of the subject **Operational Game** (Fig.3).

Now, starting from the centre, an arrow is followed downward to **tactical game**, further to **defensive game**, and from there to the descriptor *air defence* which has been "borrowed" from another circle-scheme (scheme 52) and hence has not been framed.

If, in the same report, the subject of a **reconnaissance game** is treated this descriptor will be added also.

Scanning along the second circle the descriptor **air game** will be noted down as well. When this has been done the report has been defined by 6 descriptors.

If at a later stage a question related to this subject is received then only two of the assigned descriptors would bring this report forward. For instance the codes for **air game** and **tactical game** would suffice.

The visual display of the descriptors compels the use of the prescribed thesaurus terms at the output as well as at the input activity.

The discussion on this paper follows on page 110.

NETHERLANDS ARMED FORCES SCIENTIFIC AND TECHNICAL DOCUMENTATION AND INFORMATION CENTRE

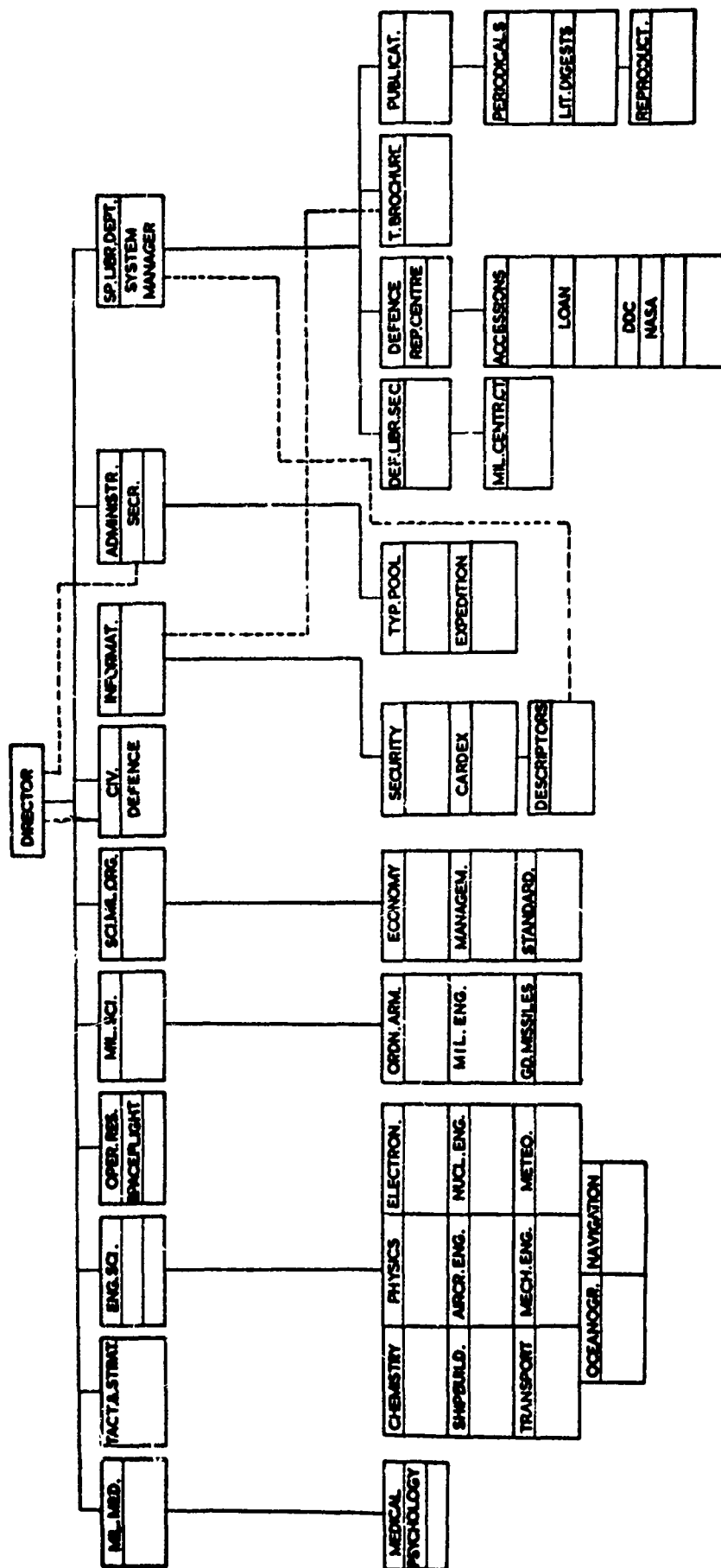
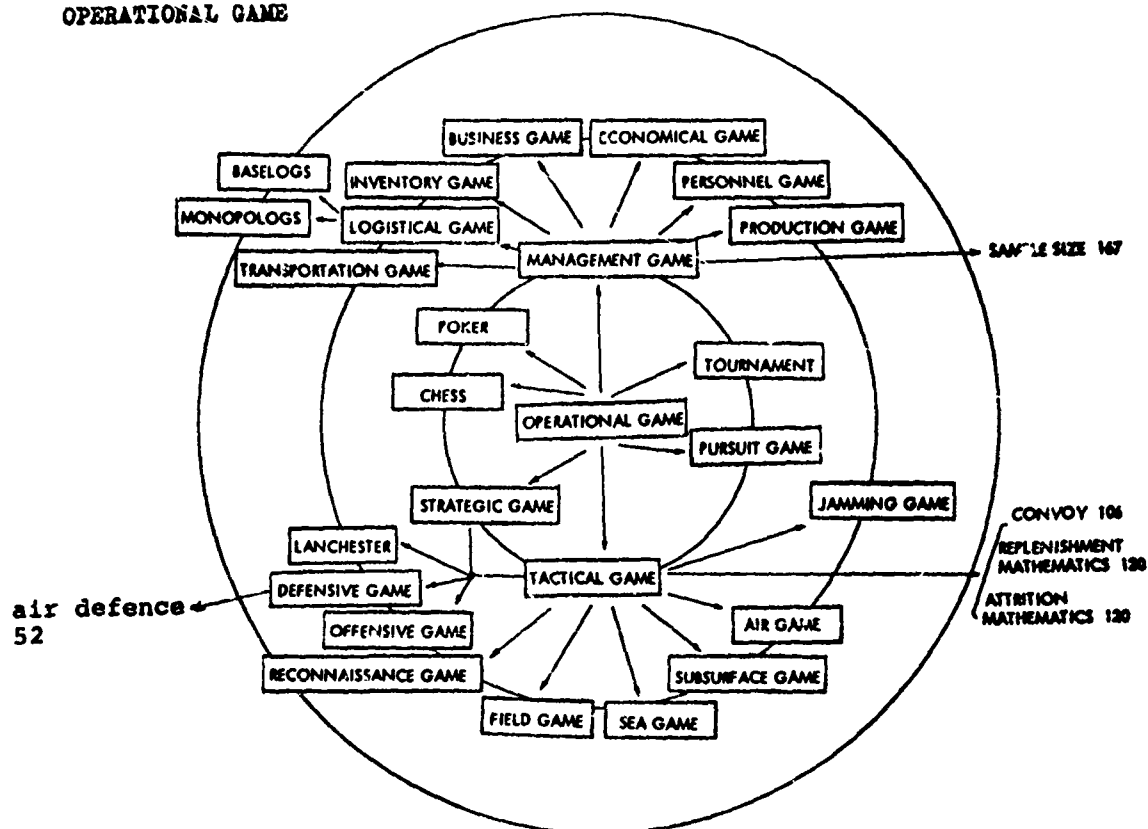


Figure 1

TDCK SOURCES INDEX		LOCATION															
CARD SYSTEMS																	
TDCK COMPACT SYSTEM	K																
UDC	K																
CODE CLAS LUCHTVAART CCL	K																
INST SCHEEPVAART EN LUCHTVAART ISL	K																
TECHNICAL BROCHURES	BROG																
COOPERATIVE AUTHORS INDEX	RC																
BOOKS OF REFERENCE																	
CODEX MEDICUS	G																
DEHEMA WERKSTOFF TABELLE	C																
ENCYCLOPEDIA CHEMICAL TECHNOLOGY KIRK - OTTMER	C																
MEDICAL DICTIONARY DORLAND'S	G																
ABSTRACT JOURNALS																	
AED-AB WFO ZUR KERNFORSCHUNG UND -TECHNIK	N																
ARMS CONTROL AND DISARMAMENT	PUBL																
ASME INDEX 1880-1988	W																
AUSTRALIAN SCIENCE INDEX	PUBL																
BATTELLE TECHNICAL REVIEW ABSTRACTS	PUBL																
BSRA ABSTRACTS	S																
BULLETIN SIGNALETIQUE SUIT	L																
CORROSION ABSTRACTS	C																
ELECTRONICS ABSTRACTS	E																
ENGINEERING INDEX 1988	W																
INDEX AERONAUTICUS	L																
INDEX MEDICUS	G																
INDEX US GOV RESEARCH REPORTS	INFO																
INTERNATIONAL ABSTRACTS ON OPERATIONS RESEARCH	OR																
INTERNATIONAL AEROSPACE ABSTRACTS	L																
METEOROLOGICAL AND GEOSTROPHICAL ABSTRACTS	M																
MONTHLY CATALOG U.S.A.	INFO																
NLL TRANSLATIONS BULLETIN	PUBL																
RES AND DEV ABSTRACTS. MIN OF TECHNOLOGY	INFO																
SCIENCE ABSTRACTS. A PHYSICS ABSTRACTS	N																
SCIENCE ABSTRACTS. B ELECTRICAL ABSTRACTS	E																
SCIENTIFIC AND TECHN AEROSPACE REPORTS INDEX NASA	RC																
SOVIET ABSTRACTS MECHANICS. MIN OF TECHNOLOGY	PUBL																
STATISTICAL THEORY AND METHODS ABSTRACTS	OR																
TECHNICAL ABSTRACTS BULLETIN DDC	RC																
TRANSACOM BULLETIN EURATOM	PUBL																
US AIR FORCE RESEARCH RESUMES	RC																
US GOV-WIDE INDEX TO FED RES AND DEV REPORTS	RC																
US NUCLEAR SCIENCE ABSTRACTS	N																
VDE SCHNELLBERICHTE	E																
WORLD INDEX OF SCIENTIFIC TRANSLATIONS	PUBL																
SELECTED RAND ABSTRACTS	RC																
ENVIRONMENTAL EFFECTS ON MAT AND EXJPM ABSTRACTS	C																

Figure 2

OPERATIONAL GAME



OPERATIONAL GAME

Operational game E 4 R
 Tactical game E 14 R
 Defensive game E 16 Q
 Reconnaissance game E 9 R
 Air game E 13 Q

Air defence 52

Air Game E 13 Q
 BaseLogs O 6 D
 Business Game E 14 Q
 Chess E 15 Q
 Defensive Game E 16 Q
 Economical Game E 17 Q
 Field Game E 18 Q
 Inventory Game E 19 Q
 Jamming Game E 20 Q
 Lanchester O 4 D
 Logistical Game E 1 R
 Management Game E 2 R
 Monopologs O 5 D
 Offensive Game E 3 R
 Operational Game E 4 R
 Personnel Game E 5 R
 Poker E 6 R
 Production Game E 7 R
 Pursuit Game E 8 R
 Reconnaissance Game E 9 R
 Sea Game E 10 R
 Strategic Game E 11 R
 Subsurface Game E 12 R
 Tactical Game E 14 R
 Tournament E 15 R
 Transportation Game E 16 R

E 13 Q
 O 6 D
 E 14 Q
 E 15 Q
 E 16 Q
 E 17 Q
 E 18 Q
 E 19 Q
 E 20 Q
 O 4 D
 E 1 R
 E 2 R
 O 5 D
 E 3 R
 E 4 R
 E 5 R
 E 6 R
 E 7 R
 E 8 R
 E 9 R
 E 10 R
 E 11 R
 E 12 R
 E 14 R
 E 15 R
 E 16 R

Figure 3

OPERATIONAL RESEARCH

100	MATHEMATICAL STATISTICS
104	OPERATIONAL RESEARCH
106	QUEUEING
108	MATHEMATICAL PROGRAMMING
110	GAME THEORY (MONTE CARLO)
112	OPERATIONAL GAME
114	EFFORT DISTRIBUTION
116	INVENTORY CONTROL
118	PLANNING
120	WARFARE MATHEMATICS
122	PROBABILITY CALCULUS
123	HITTING PROBABILITY
125	SYSTEMS ANALYSIS
129	AVIATION MATHEMATICS
135	MEDICAL OPERATIONAL RESEARCH

Figure 4

DISCUSSION

H.A.Stolk: Why does TACK use a letter-number-letter code instead of a purely numerical coding system?

J.A.Schüller: The use of this notation makes it possible for us to post a total of 12,500 descriptors on 1,000 cards.

C.O.Vernimb: What is the annual input of documents into your system?

J.A.Schüller: The input is 40,000 to 50,000 documents per year.

S.C.Schuler: (a) Do you find it practicable using your manual system, to send copies of abstracts direct to groups of scientists on an SDI basis?

(b) Do you use microfiche as a means of sending out documents? What is the reaction of users?

J.A.Schüller: (a) We do not have an SDI system but we do send information to workers on a continuing basis if we know they are interested. One disadvantage of this is that they neglect to let us know when their interest in the subject ceases.

(b) We do have many microfiche but we do not make very much use of them because of lack of good reading equipment. Users still prefer hard copy reports.

PAPER 10

MECHANICAL SYSTEMS

by

N.E.C. Isotta

**European Space Research Organization,
Paris, France**

SUMMARY

The view is put forward that the handling of large document files requires mechanization and that even processes such as document analysis for input, question analysis for retrieval and retrieval result evaluation, must eventually succumb to machine treatment. Key punching of computer input presents particular problems. The solution could be optical scanning if standardized print formats were used in document production. The direct interrogation of the machine file by remote visual display consoles is an inevitable development. The ESRO/ELDO Documentation Service hopes to have a system available in Europe early in 1969. Ultimately such consoles would be installed at strategic points throughout the European network of ESRO establishments.

MECHANICAL SYSTEMS

N.E.C. Isotta

1. INTRODUCTION

Until not very long ago, perhaps only about four or five years, the authors of most papers on mechanised systems of information retrieval, would be mainly concerned with one of two things, and these were in very general terms, either the necessity for the justification of the commencement of a machine system, or an attempt to prove that having started such a system, the results were worthwhile or at least as good as expected. Nowadays, the atmosphere in these matters is rather different since most people, and by this I mean both the customer and the supplier, have realised and accepted the need for mechanical methods of handling large files. However, there is considerable divergence of views on the level at which mechanisation really becomes necessary. Probably for report literature the figure could be as low as 100,000 items. But in actual fact this is also partly a function of the slow development of any form of any standardised vocabulary. Since scientific and technical development always produces a corresponding vocabulary, it becomes essential that existing systems should lend themselves to adaptation without a large amount of manual effort. Machine systems with built in feedback principles are obviously necessary if we are easily to keep up with developing technologies.

For non-information conscious administrations, the "subjective threshold of acceptance for a machine system" depends to a large extent on existing familiarity with large bodies of material, or large numbers of items of any kind. For example, a motor manufacturer used to large quantities of stocks of spare parts of 100,000 different items would probably not be convinced of the necessity for mechanisation for a *simple* reports field until the store could be described in terms of "nearly a quarter of a million", i.e. something over 200,000. On the other hand a manufacturer of nuts and bolts might well be convinced at "almost 100,000 items". Pressure from the potential user is rarely strong enough, or well organised enough, to affect the situation.

2. THE MACHINE VERSUS THE PROFESSIONAL

Eventually a certain amount of time is usually allocated on a computer which is primarily intended for other purposes, e.g. payroll, stock control, or as in our case, scientific data processing. It is here very often that the first troubles begin, particularly if use of the system is sufficient to demand real time operation. There is certainly complete acceptance of the fact now, that documentary processes are particularly amenable to mechanisation in what are often known as "business activity" areas. Such areas may be fairly easily defined; they include operations which can theoretically be performed without the direct intervention of professional labour, even though such labour may have been necessary initially to establish the operational procedures. These will include such matters as stock control, catalogue or index printing, preparation of announcement journals or accession lists, the establishment of "field of interest" registers etc. The most important activities remaining which still require professional attention are therefore document analysis for input, question analysis for retrieval, and retrieval result evaluation. Speaking rather heretically, primarily as a documentalist, and not as a user, it seems to me however inevitable, that even these areas must eventually succumb to machine treatment, simply because of the sheer weight of material involved, in conjunction with the increased effectiveness of the machine systems available.

3. MACHINE INPUT AND COOPERATIVE SYSTEMS

The main problem area for many years has been the one of getting the documentary material into the machine. Keyboarding in one form or another has remained essential. Technically speaking it has been possible for some time to arrange for input to be made directly to a computer without the necessity for such keyboarding operations. Economically however, such systems, which are normally accompanied by very large workload capacities, have not been justifiable in circumstances where the capacity would never be fully taken up. This is therefore still one of the most expensive, time consuming, and error ridden parts of an integrated machine system. Optical scanning could be a solution if standardised print formats were used in document production, in order to avoid an intermediate keyboarding into such a standard type script.

The ESRO/ELDO Space Documentation Service based on an exchange arrangement with NASA, is one of the first ventures of the kind where the agency receiving the machine system i.e. ESRO/ELDO, is also responsible for the provision of machine input to the system operated by the supplying agency i.e. NASA. This has underlined the problems mentioned above and has certainly indicated the enormous advantages which could be gained if greater standardisation in this area could be achieved. In spite of the difficulties involved, however, material now being processed by ESRO in Paris, according to standard NASA procedures is about to be fed, through the medium of punched paper tape, directly to the computer system responsible for the Photon production of the NASA STAR journal. I may say, that there is great satisfaction in both the NASA and ESRO centres at the successful outcome of this operation, which has been made possible only by great patience and understanding on both sides. As part of the exchange arrangement NASA has generously made available to ESRO/ELDO its total machine system together with the relevant file of information on magnetic tape. In addition, microfiche of a large number of the items quoted on the file are also provided. The service thus provided by ESRO/ELDO is available to both ESRO/ELDO staffs and to authorised users in Member States, and members of Eurospace.

4. MACHINE OUTPUT AND USER REACTION

It is clear that in spite of the advanced computer age in which we live, there has been a general diminution of standards of production resulting from the use of computers, and many of the users of machine document systems are accepting this with reluctance. The computer manufacturer's philosophy until quite recently has been that the advantages inherent in machine processing in respect of time saving, and capacity, have outweighed any disadvantages apparent in the final machine product. In my view, they have been totally wrong. The manufacturers of such things as detergents can teach the computer manufacturer a great deal concerning "eye appeal" and "packaging". There are known cases where the cost of the package is greater than the cost of the contents; one specific example outside the detergent field, is the can of water supplied on certain European flights. Even now upper and lower case computer output is a rarity and is often associated with some other extremely expensive off-line printing machine. However, by now, the user too should have become somewhat more sophisticated in his reaction to the current standards of computer printout. He should make the best of what is available since it is a retrograde step to interpose between the computer output and the users, some intermediate manual stage, be it editing or the improvement of appearance of the output by some other printing or reproduction process. I feel sure that what must be aimed at, is a completely satisfactory direct computer output; but certainly, in the meantime, the user must overcome his prejudices, although at the same time he should be sufficiently vocal to indicate that the result is not really pretty enough to encourage him to make the greatest use of it.

The question arises as to the best method of placing the user in contact with the body of information available. In our case, apart of course from our own staffs, the contact is through the medium of correspondence and telephonic communication with the documentalist who is to pose the question to the computer. Contact is thus to a large extent remote. Our experience with our own staff shows that in this respect it is difficult to match the

results achieved by personal interview between the user and the documentalist. Is there therefore a substitute for such personal contact? Almost certainly the answer must be direct interviews with the computer itself.

5. DIRECT USER ACCESS TO THE MACHINE

I have no doubt that the Orwell 1984 concept of a "Big Brother" machine is quite possible within the time available between now and then. Such a machine would almost certainly be capable of a wide variety of jobs, medical diagnosis being but one example which springs to mind. Such operations, however, could only be carried out on a governmental basis (hence the Orwell concept) with users subscribing to the terminal equipment just as they now do for their telephones. From an individual organisation's point of view however, there could be distinct advantages in having smaller, cheaper private machines - and there would also certainly be a commercial interest for the computer manufacturer in providing such machines. In the end of course, someone will also consider what the user himself would like.

A habit which is, I think, engrained in most of us after centuries of the existence of libraries, is that of browsing. This is something that the machine has been tending to deprive us of, since somehow, wading through a computer listing is not quite the same thing as browsing through a shelf of books. It is now possible however, to approach a similar situation by means of direct interrogation of the machine file using a remote visual display console. The ESRO/ELDO Space Documentation Service hopes to have such a capacity available initially for its own analytical staff, early in 1969, closely following a NASA lead. Ultimately such consoles would be installed at strategic points throughout the European network of ESRO establishments, thus enabling the user to go direct to the machine as and when he feels like it. For some time it has, I think, been apparent that future development would be in this direction. It is essential that the future of machine information retrieval is not designed around the capabilities of the first and second generation computers with which the technique was born. Joint effort on the part of the supplier, i.e. the documentalist, and on the part of the user should soon achieve the desired result.

DISCUSSION

H.A. Stolk: What services does ESRO documentation unit provide and to whom is it provided?

N.E.C. Isotta: The unit can search back in files dating to 1962, covering 300,000 or more references, in subject searches. It provides an SDI service on individually constructed profiles but intends to transfer to standard profiles soon as this gives a very much cheaper service. The service is provided to ESRO and ELDO staff, members of Eurospace and to authorised users in member states.

J.R.C. Licklider: I cannot understand how you get "immediate" indication in a mechanized system working in the conversational mode. Take the example that you have 10^6 documents and 10^3 descriptors, and that a typical retrieval attempt is specified by 6 descriptors. Also assume thirty users in a multi-access interactive system. If you stored with every pattern of six descriptors, the number of patterns associated with it, there would be about 10^{18} items in the file and that would not be reasonable. If you stored with each descriptor the identification of all the documents associated with it (1,000, or perhaps 10,000 or 100,000) you would have to transfer data from a slow secondary to a fast primary memory six times and then evaluate the Boolean expression. The waiting time would be 15 to 30 seconds. Is the key to limit the size of the file to say, 100 items?

N.E.C. Isotta: I cannot explain how the system works but I have seen it working at Lockheed Corporation in San Francisco and at the NASA facility in Washington.

C.D. Vernimb: Judging by experience of rejection by users, how many irrelevant documents are they prepared to accept in the results of a search?

N.E.C. Isotta: Users vary tremendously over the amount of material they are prepared to look at. It is difficult to be very definite on standards of precision as this depends too much on the individual user.

PAPER 11

AN INTRODUCTION TO THE STUDY OF
COST EFFECTIVENESS IN INFORMATION SYSTEMS

by

Professor J.N.Wolfe

Edinburgh University, UK

SUMMARY

Observations on the nature of cost effectiveness studies in general are made as an introduction to the procedures being adopted in a study of information services commissioned by the Office of Scientific and Technical Information, UK. Cost determination for alternative types of service is the first step in the procedure. The replacement of an old information service by a new type and the situations in which two alternative types of service exist side by side are evaluated. Finally, the services provided by alternative information systems must be evaluated.

AN INTRODUCTION TO THE STUDY OF COST EFFECTIVENESS IN INFORMATION SYSTEMS

Professor J.N. Wolfe

1. THE NEED FOR COST EFFECTIVENESS STUDIES

Large sums are spent each year on information services in each of the NATO countries. As an example, the United Kingdom alone spends about 50 million pounds each year on library services only. We have now no reliable and consistent statistics for the amount spent in other NATO countries, and in particular, we lack information on the amount spent on information services other than libraries. The OECD is in the process of attempting to collect this information and the study is underway under the direction of the Studiengruppe of Heidelberg, Germany.

We know, however, that the total sum being spent is sufficiently large and growing with sufficient rapidity to present a serious economic problem. This economic problem has several aspects. First, there is the question of how much ought to be spent on information services in general. Secondly, there is the question of how rapidly this sum should grow. Thirdly, there is the question of the most appropriate division of expenditure among the competing types of information service which might be offered, and fourthly there is the question of the most appropriate organisation of information services both within a single country and between countries.

These sorts of questions may have seemed to be of only academic interest during the last decade or so, for there has been general agreement that the volume of funds available for information services has hitherto been too low, and funds have been expanded with considerable rapidity. During this period too there has been rapid technological change in the information industry. There are now many more technically developed candidates for absorption of information funds than was the case even a decade ago. As new techniques pass from the laboratory and pilot stage into the world of practical possibility the question of economic viability and value for money becomes a very real and pressing one.

2. THE OSTI-OECD STUDY OF ECONOMICS OF INFORMATION SYSTEMS

It was in this context that the Office for Scientific and Technical Information in the United Kingdom, acting in collaboration with the OECD, decided to undertake a study of the economic aspects of informatic systems. The study was commissioned from the Department of Economics in the University of Edinburgh, and involves a large team of workers including five full-time economists and a full-time information officer, two accountants, two statisticians, and five part-time economists. The work has been underway for approximately four months but will not be completed until the end of the calendar year 1969. One aspect of this work which is already rather far advanced is an economic study of the library system and particularly the public library system in the United Kingdom. It is proposed to publish very shortly a volume of essays on this topic. Most of the papers involved are quantitative and econometric in character and it would be difficult to summarise any of them briefly. I would however like to mention here two papers in particular which seem to me to offer considerable interest. One of these is a paper by Mr. Ralph Young on the Forecasting of the Demand for Library Services in the Public Library Sector by econometric means. This paper provides what I think is the first attempt to offer a quantitative forecasting technique for library demand which is not simply an extrapolation of past trends.

Mr. Young shows that even at this early stage of analysis it is possible to forecast the appropriate level of library provision in a general way at least with considerably improved accuracy. This technique has been applied, as I say, to the public library system but I think that it offers considerable possibilities of extension to library systems within private firms or government agencies. Another paper of some interest is that prepared by Dr. Jacob Moreh which examines in a statistical and econometric way the problem of economies of scale in library services. Dr. Moreh attempts, and I think for the first time, to go below the level of simply comparing large groups of dissimilar libraries with one another on the basis of an average cost figure. Such a procedure, while common enough in practice, is of course statistically exceedingly unreliable. Dr. Moreh on the other hand utilises techniques made familiar in production function studies to examine the cost functions of operating within the public library system on the basis of a variety of independent variables including number of branches in each library system, the number of employees, the number of volumes, and the volume of ancillary services such as gramophone record issues. While his results are not yet completely analysed, they do seem to indicate that the popular belief in economies of scale in the library world may be somewhat over-simplified.

3. THE NATURE OF COST EFFECTIVENESS STUDIES

Before moving to some account of the larger economic study now underway, it may be useful to provide some introductory observations on the nature of cost effectiveness studies in the context of information and library services. It will be recalled that cost effectiveness techniques were given substantial development by work undertaken on behalf of the United States Department of Defense largely in the Rand Corporation of Santa Monica, California. Put in the simplest way, the notion of a cost effectiveness study is an attempt to discover the relative magnitude of costs and benefits accruing from alternative forms of expenditure. More concretely, the early studies involved assessment of the relative cost per ton of bomb delivery for example. The essence of a cost effectiveness study is the reduction of the benefits of alternative task systems to some kind of commensurable unit. Once this is done the problem becomes merely one of comparing the alternative task outputs with their costs.

Looking at the matter in another way, we may see the cost effectiveness study as simply an improvement on the more normal cost study. The traditional cost procedure involves an examination of the costs of two alternative tasks. But clearly costs are not a sufficient determination of which task provides the best outcome.

We must consider as well the benefits achieved in each outcome.

Let us take an extremely simple example drawn from everyday life. Supposing we wished to determine which was the wiser purchase, an orange or a lemon. We could easily determine the cost of the orange and the cost of the lemon. The question of which of the two fruits provides the better buy for money depends however upon what we wish the fruits for. If we are anxious to obtain a given quantity of Vitamin C, for example, it may well be that the lemon provides the better bargain. If our object is to provide a refreshing morning drink, and we wish therefore to maximise the sugar content of the citric juices, then a different answer may be obtained. We cannot therefore tell which fruit it would be worth our while to purchase until we determine the objectives for which we are purchasing them.

4. ESTABLISHING THE COST OF INFORMATION SERVICES

With this introduction in mind there should be little difficulty in understanding the procedure which is being adopted with respect to cost effectiveness in information services. The first part of our job is to determine cost for alternative types of service. This presents certain features of difficulty because of the fact that information services, like most public services, do not normally keep accounts upon what is called a functional basis.

That is to say, the accounts of most information services take the form of a list of expenditure by name item of expenditure. That is to say labour, materials, rent, heat, etc. They do not normally assign these expenditures to the manifold functions which an information service in fact attempts to achieve. It is therefore necessary to recast the accounts of the information services in functional form before any serious further work can be done.

One of the basic difficulties here is of course the assignment of overhead costs to the various alternative functions. We have to ask for example what proportion of the time of a head librarian ought to be attributed to his work as head of an information service as well as of a library service, in a unit which offers both library and information services. Similarly we may ask what proportion of the cost of heating an information centre is to be attributed, let us say, to the preparation of abstracts on the one hand or to the preparation of translations on the other. It will be clear that, however much care is taken, there will be a certain measure of arbitrariness in such calculations. It is our object not to eliminate arbitrariness entirely, but rather to reduce it to manageable proportions.

One important issue is the extent to which information services may be added to existing library activities at lower costs than information services can be provided, in a purpose-built organisation. On the one hand we might expect that the sharing of certain overheads with a library would produce lower costs in the integrated operation. On the other hand the greater expertise which can be developed in a specialised and purpose-built organisation may conceivably offer economies of substantial importance. This balance between economies of scale and economies of specialisation is, as everywhere else in industry, an important question deserving the most careful examination.

The central core of our method consists of evaluating two particular types of situation. The first is a situation in which an old type of information service is to be superseded by a new type. This situation provides alternative information on costs and also provides information on the change in value of the service received by changing over between the two systems. An alternative approach consists of examining situations in which two alternative types of information service exist side by side. For example, we may have certain organisations which utilise an advanced information service while other organisations utilise an older style of information service. Here costs and effectiveness may be compared on a cross-section basis. It will be understood, however, that in this case there may be expected to be a substantial amount of extraneous information introduced because of the possibility of underlying quality difference between the units using the technically advanced information service and those using the technically less advanced information service. The final part of our work consists in evaluating the services provided by alternative information systems. This is clearly the most difficult part of our job. It is difficult partly because previous attempts to deal with user requirements and user needs have not been directed specifically to economic investigations. There is a fundamental difference between technological criteria of efficiency in this context and criteria of economic efficiency. Ideally, one would like to obtain estimates of the impact of the information service on the productivity of the workers receiving the information service. In practice this level of productivity is likely to be very much influenced by extraneous factors. This is a particularly damaging point if we are dealing with cross-section studies of a particular industry which has different information services in different firms. We are likely, I think, to find that good information services are in fact characteristic of technologically advanced firms, and if this is the case any attempt to correlate efficiency with information services is likely to give us too optimistic a result. When we deal with changes in information services affecting all the units in an industry, we have, I think, a rather more practical proposition, although here we will, I am afraid, be hampered for some time yet by a shortage of instances. I would expect, however, that as the number of information services examined increases, a statistically reliable result may eventually be approximated.

There are alternative methods of obtaining effectiveness measurements from information services. Some of these consist of sampling opinion about efficiency. Others consist of obtaining objective characteristics of the functioning of the information service. But this particular problem requires further consideration.

DISCUSSION

H.F.Vessey: Have you considered the cost of not providing information when making your evaluations of system effectiveness?

J.N.Wolfe: It is not possible to make allowance for a factor of this sort. All evaluation must be based on objective data. The first attempts at quantification of a service may not give a satisfactory result but by repeated efforts it is possible to develop a satisfactory method of measuring effectiveness.

N.E.C.Isotta: The provision of information to scientists and engineers must be considered as part of their continuing education and as such its value cannot be quantified immediately. The value of a piece of information might not emerge for several years. I do not agree that the amount of information available should be considered as uniform, one of your basic premises. It is precisely the non-uniformity which we have to overcome.

J.N.Wolfe: I would certainly agree with your first point, but on a matter of obtaining administrative support for expenditure on a system it is necessary to show that it will be of some practical value.

PAPER 12

TECHNICAL INFORMATION SERVICES
AND USER NEEDS

by

W.C. Christensen

Department of Defense, U.S.A.

SUMMARY

Three audiences for technical information are defined: the general audience, the mission audience, and the technical management audience. The information needs of these three groups are discussed. The provision of technical information by the U.S. Department of Defense is outlined, and in particular, some of the functions of the Defense Documentation Center are described.

TECHNICAL INFORMATION SERVICES AND USER NEEDS

W.C. Christensen

1. INTRODUCTION

To begin, I would like to define technical information in a way which I have found convenient. The view which I have adopted is that technical information is the generic term embracing the full spectrum of information generated or used by personnel working in the scientific or engineering domain. Technical information can then be divided into two subcategories - scientific information and technical, or if you prefer, engineering data. Scientific information is defined as technical information which adds to the general body of knowledge about a natural phenomenon, material property, or about a scientific or engineering discipline. Scientific information does not disclose a specific connection with nor application to the design, production, operation, or maintenance of an item of equipment.

Technical data, on the other hand, is technical information obtained from the design, development, manufacture, operation, maintenance, and logistic activities and is used by the recipient to design, produce, operate or maintain equipment. For example, technical data includes design data, development data, production data, manufacturing data, logistics data, and maintenance data. This distinction between scientific information and technical data is important since, as will be shown later, our major technical information problems are associated with technical data - not scientific information.

Now that we have established some boundaries on the subject we are dealing with, let's take a look at the general categories of audiences who use technical information.

2. WHO USES TECHNICAL INFORMATION?

As shown in Fig.1, there are three major audiences for technical information - the general audience, the mission audience and the technical management audience.

Technical information used by the general audience is characterized by the fact that the generator of the information does not know who specifically will use the information or when. As an example, we have over 850,000 U.S. Department of Defense technical reports centrally stored and available from the Defense Documentation Center. Most of these reports were required to document the results of Defense research and development efforts. However, the secondary use of this information by the general audience may be for purposes totally different from those for which the work was undertaken and at a time considerably removed from that during which the information was generated. This diversity of uses and time differential creates serious problems in effectively retrieving and employing the information. This retrieval problem is growing more difficult as the degree of technological sophistication increases. Our primary difficulty is that the technical documents are written in relation to a specific end goal which was the basic objective of the work. Many times this end goal involves a highly complex piece of equipment such as a missile or a tank which involves a multitude of discrete innovations all of which are combined to produce the end goal. The degree to which each discrete innovation is documented is highly dependent on the importance attached by the generator in relation to the end goal. This creates two difficulties. First is the ability to index each discrete piece of technology so that the report can be retrieved when a user requests the information.

The operator of the information storage and retrieval system is faced with the classic dilemma - if he employs a large number of characterizing or search terms - searches will produce a great number of documents, many of which are not particularly relevant to the user's needs. On the other hand, too few terms will result in many relevant documents going unidentified.

The second difficulty resulting from end goal oriented technical reports is that frequently there is not enough information related to a specific technology for the general audience user to effectively take advantage of the past work.

I will go into more detail on the trials and tribulations of information storage and retrieval later but for the moment, let's turn our attention to the mission audience. This audience is characterized by a close coupling to the generator of the information. The mission audience could be the procurement organization for a new piece of military hardware. In this case the research and development people are well attuned to the information needs of the procurement people with the results that not only is the precise information needed displayed, but it is also displayed in a manner most meaningful to the user. This close coupling between the generator and user results in efficient information transfer. However, we often find that the information tends to stay within the relatively narrow confines of the generator-mission user environment even though it could be of considerable use to the general audience or other mission audiences.

Finally, we have the technical management audience which I have represented by the classical pyramid. The increasing expenditures for research and development along with the additional complexity of the efforts themselves have increased emphasis on timely and accurate technical management information systems. Within the U.S. Department of Defense we have been developing a very sophisticated technical management information system covering our numerous research and technology efforts. This automated system is designed to tell users what work is being done, by whom and in very abbreviated form what the progress is. While the system was primarily designed to meet a management need, we have found that over half of the users are working engineers and scientists. These people use the system to identify on-going research and technology efforts related to their particular areas of interest. While the technical information content is minimal, it is normally sufficient to determine whether the performer should be contacted for detailed information.

3. USER NEEDS

Within these terms of reference, we in the Department of Defense have been very concerned with what technical information does the user really need and how well are our various technical information services fulfilling these needs?

To obtain at least a partial answer to this complex question, we have run two comprehensive user needs studies - one concerned with the needs of engineers and scientists employed directly by the Department of Defense and another covering those associated with Department of Defense contractors.

I have summarized the results of these two studies, performed by two different contractors, in Fig. 2. I want to go through these in some detail because the information is quite revealing in terms of our present information services and what we should be striving for in the future.

Before discussing the various information gathering characteristics of the users, a few words on the characteristics of the users themselves are in order. First, most of the technical information users are engineers or are working in engineering related areas. Too often this point is overlooked and equal or greater attention is given to the scientist and his information problems. While I do not want to belittle the information problems of the scientists, it is engineers and other applicers of technology which are my chief concern and unfortunately, their information problems are exceedingly complex.

Now let's look at how our users obtain information and the type of information they need. The first statistic pertains to the desire for information in a short period of time. While our work showed that over 20% of the users needed information in less than one day, most users would really like to have their information needs met instantaneously. What frequently happens is that the user makes a quick minimum effort at getting information. If the optimum information is not found during this first try, he will too often resort to the use of readily available but less than optimum information. For example, an engineer selecting materials may not use a low cost material because he cannot readily determine its characteristics in a particular environment. Instead, he picks an expensive alloy which he knows will do the job. This gives rise to one of the frequently used arguments against expending resources to provide better technical information systems - the users seem to do their job without them! However, the real question is, "How much could their performance be improved by instituting better technical information systems?"

The next item pertains to how the user gets his information. Our studies show that he turns to a colleague or his personal files as a first source of information which supports my argument that users operate on a minimum effort principal as far as requisition of technical information is concerned.

The next item is very important from a user need point of view. As I mentioned previously, our main concern should be with the engineer or technologist and this statistic clearly bears out the need for so called engineering type information. Yet this information consisting of design information, test data, operational data, manufacturer's part and component information and the like, is the most difficult to handle in a technical information system. One aspect of the problem is that engineering information is difficult to capture so that it can be incorporated in an information system. The difficulty stems from both the amount of information being generated and the fact that most of it is being created for the mission audience which is not particularly motivated to disseminate it to the general audience. However, the more serious problem with engineering information is that it tends to have a short half life. In other words what may be valid up to date engineering information today may be obsolete tomorrow. We have run some experiments with user oriented information systems where we have incorporated both engineering information which users knew was up to date and some engineering information which the users were not quite sure of. The results were that the engineering information which the users were not sure was up to date wasn't used at all - even though it was probably better information than they could obtain from other sources.

The cost of maintaining quality control over short half life engineering information is very high. For instance, I estimate that the U.S. Department of Defense expends about \$80M a year just to operate its military specifications and standards program. When one begins to consider expanding this type of quality control to other sources of engineering information, serious questions of the cost versus benefits must be raised.

Along this line I would like to mention one of my "pet" concerns about the information utilization habits of engineers - recalling that the engineer normally obtains his technical information from his local environment, that is, his personal files and colleagues, take a look at his private library sometime. My experience has been that his favourite tools are often text and reference books obtained in college plus a few odds and ends he has encountered and used in depth during his career. His college books contain information generated at least five years before the book was published. Adding this five years to the time since his graduation means that the information is on the average 20 to 30 years old. The various odds and ends that he has picked up over the years are similarly in various stages of obsolescence. To me it is a wonder that he can survive in this age of exploding technology and multi-disciplinary efforts.

4. PROVISION OF TECHNICAL INFORMATION BY DEPARTMENT OF DEFENSE

Now that we have addressed the user needs, albeit in abbreviated fashion, let us take a look at the existing U.S. Department of Defense's situation from a technical information system viewpoint.

First, we have a large central depository for technical reports resulting from Defense research and development known as the Defense Documentation Center. The Center accessions about 50,000 new technical reports every year. The reports are indexed upon receipt and the bibliographic information added to a computer based search system, and at the same time, announced to the defence user community. Subsequently, the reports can be ordered or a bibliography can be prepared on any given subject. At the present time, the Center is receiving about 2 million request for technical reports and 20,000 requests for bibliographies each year. Granted the use factors are impressive, but some consideration must be given to the Center's operation in terms of the user needs. First, considering the input side of the Center, we have a major problem which I mentioned in the beginning. The technical reports handled by the Center are prepared in relation to a specific end goal. While these reports may have a high degree of relevancy to those intimately concerned with that particular end goal, their effectiveness as information transfer media to users not familiar with the area of endeavour, may be low. We also have the problem that too often the actual technical information content of these reports is low and as the saying goes "garbage in - garbage out". The real problems become visible when attempts are made to characterize the contents of these reports. It would be fine if the users only needed to retrieve information on an end goal basis such as development of solid propellant missiles. However, more and more we find that users are searching for discrete pieces of technology associated with a particular problem at hand such as pressure sealing of gauges. Now this information might be reported in a missile development report if it was particularly pertinent to the overall missile development programme. The problem is that indexing the report so that each discrete piece of technology is reported, results in a large data bank which is difficult to effectively search and more importantly, results in an unacceptable large number of irrelevant report identifications in response to a user's query. Nothing can discourage a user more than loading him up with a vast amount of information which he is not interested in. There is one further problem associated with the operation of a central report depository such as the Defense Documentation Center. This is the time delay associated with obtaining the information. Regardless of how efficient the Center's operation is, there is about a 2 week delay primarily as a result of physical transfer of the request and resultant product. The importance of this delay can be seen when the users desire for rapid access to information is considered. There are two ways to get around this situation - utilization of advanced communication techniques or to provide the information in advance to an information centre in the users immediate environment. Several other speakers are covering advanced communication techniques so I will not dwell on it here, except to mention that we are installing several experimental remote on-line terminals to the Defense Documentation Center.

The providing of technical information to the user locally has been the traditional role of the technical library. The difficulties are many fold. To begin with, they deal in documents - not information. The user must research the documents and extract that information which is pertinent to his needs. Also, the technical libraries find it increasingly difficult to maintain collections covering the full range of the interests of the users they service. Finally, there is a communication problem between the technical user and the non-technical librarian. I feel that this latter point is particularly significant and that if our so called retain stores are to become a viable part of our technical information systems of the future, they must employ technically competent personnel in addition to those solely concerned with storage and retrieval of documents. These technically competent personnel which we might call technical information specialists not only provide an effective coupling between the user and the information source, but can also answer users' queries with highly relevant information - not just documents.

One area where the U.S. Department of Defense has created technical information systems manned by technically competent personnel is the information analysis centres. From a technical information transfer point of view, these centres are very effective. Each of our 26 centres is assigned a very specific subject or discipline area. Generally, the personnel operating the centres actually spend a portion of their time working in the subject or discipline area, providing specific answers to users' inquiries in their area of expertise and publishing high technical content documents. Thus, these centres get around

the input problems associated with the operation of central document systems like the Defense Documentation Center and provide the personalized coupling between the user and the information. Because of their competence in their field of expertise, they also provide the quality assurance factor which I discussed in relation to engineering information. The major drawback to these centres is that they are very expensive and to date, we have only been able to justify them for a limited number of subjects or disciplines.

This brings me to a key point which we must face in the technical information business. The cost of various technical information systems and services can be identified. However, the benefits in quantifiable terms are very difficult to ascertain. Intuitive arguments that technical information systems and services are good have just about exhausted their appeal. Within the U.S. Department of Defense we are initiating a program of charges for selected technical information services on the basis that if the service is of value to the user, he should be willing to pay for it.

In summary, I see three pressing needs for technical information systems of the future. First, we must improve the quality of the technical information in our systems and I would suggest that the best place to do this is at the source. Next, we must get more users involved in the design and operation of technical information systems. Too often, systems are created to serve phantom audiences. Finally, we must find ways to quantify the benefits derived from more effective technical information systems so that decisions to establish these systems can be based on fact and not fantasy.

The discussion on this paper follows on page 131.

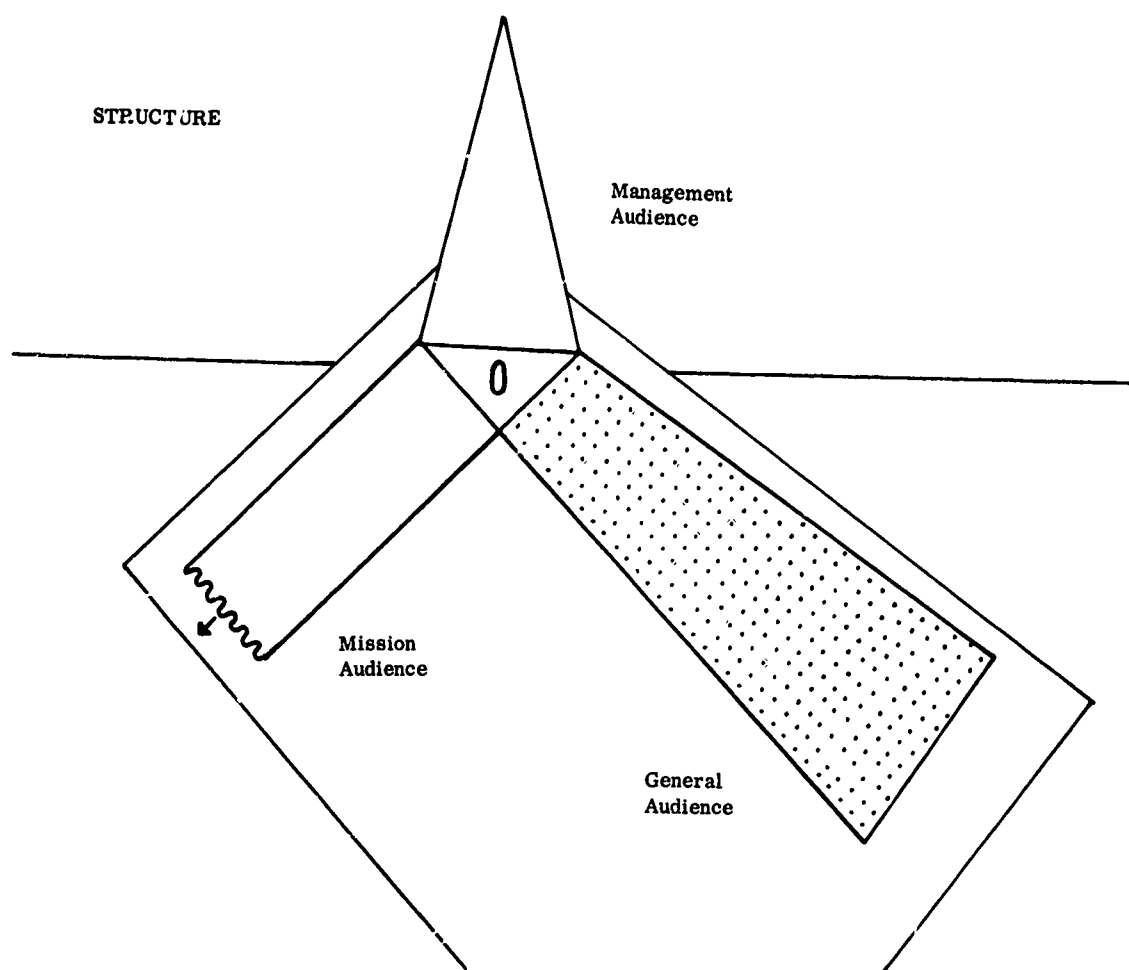


Figure 1

<u>SEARCH AND ACQUISITION</u>	<u>U.S. DoD Users</u>	<u>U.S. Def./Ind. Users</u>
o Desired information in less than 1 day	21 percent	23 percent
o First source for information		
Mode	Colleague	Colleague
Local work environment	60 percent	51 percent
Use of Libraries	5 percent	10 percent
Use of DoD information systems	0.04 percent	1.30 percent
o Use of engineering information	42 percent	40 percent

Figure 2

DISCUSSION

R.W.G. Gandy: What steps would you propose to improve the quality of information at the source, i.e., the standard of report writing and presentation, and in particular the elimination of "garbage"?

W.C. Christensen: The first step is to improve the education of engineers and scientists with regard to report writing. Secondly, the majority of reports are produced under contract because many contractors feel that the value of their work is judged by the number of reports produced. Contract monitoring agencies should encourage the state of affairs where a report is only generated to give some really useful information on the work being undertaken.

E. Keonjian: (1) What is done to reduce the amount of useless information entered into the DDC system? (2) Could you define the steps in answering an enquiry?

W.C. Christensen: (1) Technical monitors of a contract are asked to suppress progress reports produced solely on a time basis e.g., every three months. (2) Taking as an example an engineer wanting a bibliography on a special subject, the steps are:-

- (a) Question is analyzed and descriptors allocated from DDC Thesaurus.
- (b) Computer staff put request to the UNIVAC 1107 system and identify the relevant reports.
- (c) Staff with some knowledge of the requester's speciality examine the print-out and edit it, perhaps to reduce the number of references. If the number of references is very great, the requester will be asked for further definition of the subject.
- (d) When the subject specialist is satisfied, the list of references is sent to the requester.

R.D. Kerr-Waller: (1) The number of data banks needed to come with the volume of literature handled by DDC must be considerable. How does an on-line terminal system operate when the question can fall into any one of several data banks? (2) What charges does DDC propose to make for its services?

W.C. Christensen: (1) About six data banks are used, but only one is the report data bank which contains bibliographic details of about 400,000 references. A change soon to a UNIVAC 1108 system will make searching more rapid. (2) DDC have established a charge of 3 dollars for each hard copy report supplied; microfiche are supplied free. The twenty-six information analysis centers operate in various ways tailored to the needs of their circle of users. The total budget of each center has been reduced making it necessary for them to introduce charges, but each center decides for itself exactly how these should be levied.

F. Hangsted: The "General Audience" often includes the decision-makers. Their task is often made more difficult by the amount of jargon and special terminology used in the literature. Is there a solution to this problem?

W.C. Christensen: The basic solution is to get colleges to give better training in written expression. In particular cases, persuasion or some direct action can often help.

A.H. Holloway: Is there a solution to the problem that what may be essential information to some users may be "garbage" to others?

W.C. Christensen: By "garbage" I mean lengthy passages of text containing very few facts. Elimination of this style of writing would be of advantage to everyone. I appreciate that when writing a report it may be difficult to judge exactly what will be of interest to a particular audience but we must try to increase the proportion of technical content of reports. We must also make better use of reports by finding ways of identifying discrete pieces of information which may be of use outside the main field of the report.

PAPER 13

SELECTIVE DISSEMINATION OF INFORMATION

by

M.S. Day

National Aeronautics and Space Administration, U.S.A.

SUMMARY

Selective Dissemination of Information (SDI) provides individual scientists and engineers with announcements of a limited number of documents specifically of interest to them, in contrast to the general coverage provided by increasingly bulky abstract journals. Selection is done by a computer program, which compares a file of bibliographic data on current reports and journal literature with an SDI user's interest profile, then prints out references to matching documents. The selected references may be presented to the user on cards suitable for filing or on less expensive printed lists, and may provide only the document citation or the full abstract. Feedback by the user on the relevance of the documents helps to optimize his interest profile for best selection. Comparison of numerous individual interest profiles is expensive in computer time, and profile improvement requires assistance by vocabulary specialists. Economical service to large numbers of participants may be provided by the use of standard subject profiles, as typified by the NASA/SCAN (Selected Current Aerospace Notices) program which is described.

SELECTIVE DISSEMINATION OF INFORMATION

M.S. Day

1. INTRODUCTION

As I am addressing working scientists and engineers, there seems no need to belabour the trite expression, the "information explosion". I am sure that you have already encountered the problem in the shape of great numbers of reports and articles to read and digest and in the growing bulk of abstract journals that you must use to keep currently aware of developments in even a limited area of your interests. While we in the profession of information science and technology cannot say that we have kept ahead of the problem by offering fully satisfactory solutions, I feel that advances are being made. One approach to the problem of reducing your literature review efforts is Selective Dissemination of Information, or SDI for short. SDI applies the advances in computer sciences already discussed at this symposium to the task of providing a personalized current awareness service.

Current awareness services are not new. Most libraries have long provided patrons with copies of current documents that the librarian has decided will be of interest to particular individuals. Library accession lists are frequently categorized to call newly received documents to the awareness of groups of potential users. Current issues of abstract journals, when routed to pre-established distribution lists, also are current awareness tools.

But the usefulness of such methods is limited by inconsistent selection or by excessive volume of material announced. In the case of abstract journals, even a categorization scheme does not overcome the problem of scanning a great bulk of abstracts. Nor does such journal categorization, with announcement of a document only in a single category, provide for complex interests, which often cut across many fields.

Some scientists and engineers may claim that they have no need for a selective current awareness service. They may be those active leaders in their specialty who belong to the so-called "invisible colleges". Besides attending all pertinent conferences, they exchange and file preprints and reprints. For the great majority of scientists and engineers, however, a more formal and efficient service is necessary to alert them to current documents of specific significance. Even members of invisible colleges find that a current awareness service alerts them to timely reports and journal articles that might otherwise be delayed in reaching their attention.

In describing what SDI can do for you, it is also essential to refer to the mechanism of the SDI operations and especially to the relative costs and efforts involved in the many different SDI systems that can be designed. As users, you will be concerned with obtaining the best design and operation possible. Obviously, if the managers of your firm or professional society feel that a proposed SDI system is excessively costly in relation to the organization's many other goals, it will not be established. As potential users, you should be prepared to participate actively in the design of information systems and be able to demonstrate the cost effectiveness of the service you will receive.

2. PRINCIPLES OF SDI

SDI is a current awareness tool and results in the selection and announcement of current documents having a high probability of interest to the individual user. The fundamental

element in selection is the comparison, by computer, of two data files (Fig.1). One is the file of bibliographic data assigned to newly received reports. These data include subject index terms and other document representations -- the authors, corporate sources, supporting agency, contract or grant number, etc. The other file contains the users' interest profiles, which are equivalent to bibliographic search strategies (Fig.2). The interest profiles consist of bibliographic data elements, such as subject index terms, related by the common Boolean logic expressions, such as AND, OR, or NOT. Other methods of relating terms are possible; e.g., by assigning relative weights to each term, a certain minimum total weight of matching index terms then being required before a document is chosen for announcement. Authors, contracts, and other document identifiers may also be included in the search strategy. The particular features of the profile structure and the flexibility in the document identifiers that the profile can incorporate depend on the computer capabilities available.

The interest profile is of first importance in the success of an SDI system. Your profile is not just a paragraph describing your interests, it is a rational set of specific terms in the same technical language used by the document indexers. Structuring an interest profile may require considerable skill. If you were interested, for example, in the subject of supersonic transports, it would not be sufficient to put just the term "Supersonic Transport" in your profile. Documents specifically on the Concorde might be indexed to the term *Concorde Aircraft* and not to the general term *Supersonic Transport*. What other aspects of supersonic flight are you interested in -- clear air turbulence, sonic boom, general concerns of international law affecting civil aviation, or basic engineering problems involving supersonic heat transfer, supersonic flutter, or supersonic wind tunnels? Are you interested in getting every report on a given contract? Do you want to limit the number of announcements that you receive, remembering that the total number of documents indexed by certain common terms might be quite large? How is this limitation on number of announcements to be done on your interest profile -- by removing index terms, or by restricting selection through Boolean logic relationships?

As an SDI user, you would have to take an active part in structuring your profile, or else have it written for you. Because of the complexity of a satisfactory interest profile, experience with SDI systems has shown that the scientist or engineer requires considerable help in constructing his profile. Such help requires the services of a professional reference analyst, who has the authorized authority terms and indexing patterns and practices at his fingertips. I again wish to point out that the success of an SDI program is directly related to the quality of the user profiles.

3. ELEMENTS OF SDI

Besides the interest profile, features essential to any SDI program are:

- (a) A standard form for presenting selected announcements to the user. This may be in a form that the user can conveniently retain.
- (b) A method for conveniently requesting a copy of an announced document from a local library or from the central operator of the SDI service.
- (c) Routine feedback by the user to the system as to his degree of satisfaction with each document. The feedback should provide a quantitative measure of the performance of the user's interest profile and of the operation of the over-all system.

Many organizations, both in the United States and Europe, have initiated SDI programs to date. Their experience, as reported in the literature, can be drawn on in designing new current awareness systems. I am most cognizant of the programs of the National Aeronautics and Space Administration. NASA has been a leader in the SDI field, having operated several types of program since late 1963. Its SDI services have been distinguished by volume of input and by size of user population. During 1967, for example, 875 interest profiles were matched four times each month against the data files corresponding to the full contents of the current issues of *Scientific and Technical Aerospace Abstracts* and *International*

Aerospace Abstracts. Aerospace reports, journal articles, conference papers, etc., matched during the year totaled 63,700; and a total of almost 800,000 announcements were distributed. NASA is now moving into new evolutionary phases of current awareness service, as I will discuss.

4. TYPES OF ANNOUNCEMENT FORMS

Numerous forms have been designed by SDI system operators for announcing selected documents to users. A distinction can be made between the card type of announcement, with each announcement issued as a unit record, and the listing type, with the announcements printed continuously on sheets.

4.1 Card format

The majority of SDI services provide the user with a card for each selected announcement. Because SDI is generally thought of in the framework of a computerized information system, this card is typically an electronic data processing (EDP) or tab card. Systems providing edge-notched cards or other announcement form designs are feasible for information services of limited scope. The card may present a full abstract or merely a bibliographic reference. If only a document citation is presented, its limited informativeness may be enriched by also printing out the index terms assigned to the document.

The material presented on the SDI announcement card may be computer-printed, or it may be duplicated by, for example, offset printing. Offset reproduction permits full abstracts, even with special symbols and illustrations, to be reduced in size and presented on a single card, whereas a computer printout is strictly limited in the number of lines of information that can be presented. A disadvantage of offset reproduction is the need for two operational procedures. The computer first punches a card with the user's name and address and an identifying number for the selected document. The punches are then interpreted into printed characters on the face of the card. The abstract must then be reproduced by offset onto the corresponding punched card. Much handling and sorting of the cards is involved in such a dual system.

The notification cards received by SDI users are usually designed so that a stub may be detached and returned to the library for requesting a copy of the document, or for merely indicating that the announcement was or was not of interest.

Both the user's address and the abstract need not be presented on a single card, although this is the common practice. NASA's first SDI program, operational from December 1963 to January 1966, provided the user with *two* cards for each announcement (Fig.3). One was an EDP card which was punched and interpreted with the user's name and address and the document number. These cards contained small prescored blocks which the user could punch out to express his evaluation of the announcement; i.e., that the announced document was (1) of interest and that a copy was wanted, (2) was of interest but that no copy was wanted at the moment, or (3) was not of interest. The second card was not computer manipulated, although it was cut to the same size and shape as the typical computer punched card. It presented the full offset-printed abstract of the selected document. The two cards for an announcement were inserted into a single window envelope, with the user's name and address visible. As the envelopes were necessarily in order by the abstract number, they were then manually sorted according to the user's organization for batched mailing and subsequently by the organization's mail room for internal distribution (Fig.4).

Cards are very popular with the SDI user, as he may file those of particular interest in a personal desk-drawer file. Undoubtedly, this is a valuable tool for many scientists and engineers. However, maintaining an individual file, either of cards or documents, can be expensive in terms of the individual's time, and possibly in storage space. SDI is primarily a current awareness service, and provision for a continuing bibliographic data file is of subordinate value.

The cost of a particular SDI service depends on so many factors of input volume, materials used, computer processing, degree of profile assistance, geographical distribution of users, etc., that only a rough figure can be suggested for the cost of an operating system. Detailed cost analysis should precede the implementation of any SDI proposal. A card-type SDI system might fall in the range of \$100 to \$150 per user per annum for a large volume of input references; e.g., the total references in *Scientific and Technical Aerospace Reports* and *International Aerospace Abstracts*.

4.2 Listing format

Less expensive than card-type announcements are computer-printed listings of selected bibliographic references. In general, listings present only bibliographic references, perhaps with the index terms to enhance the document content information provided by the title, author, and other reference elements. If the abstracts are on machine-readable files, the abstract may be printed out in full or in part. While this may be helpful to the user's understanding of the content of the document, it incurs the expense of added computer use, increased bulkiness of the announcement package, and added review time for the user.

NASA's present SDI system, in effect since February 1966, is an adaptation of the simple listing. A three-copy, no-carbon-required form is used (Fig.5). The computer-printed bibliographic references of course appear on all three sheets, together with the user's name and address. The computer also prints blocks (lozenges) opposite each announcement. The empty blocks are for the recipient's use in checking the relevance of the announcement to him; whether it is of interest and the document is requested, of interest but the document not wanted, or of no interest. When the user receives his announcements, he marks his evaluation opposite each announcement, simultaneously marking all copies, then tears off the original for retention if he desires. The other copies are forwarded to his library, where one of the copies is used to fill document requests while the other is returned to the system operator. The operator tabulates all responses and computes the ratio of number of relevant announcements, as indicated by the user, to the total number of announcements for each user and for the over-all system. The tabulated results serve as measures of operational effectiveness. Again, costs of a list type system depend on the information presented and the other factors common to all SDI systems, but might fall in the range of 60 to 70 per cent of the cost of a card-type system.

4.3 Mixed announcement forms

Listings are adequate as announcement tools, but they lack an important element of a fully automated system; namely, machine readability of document requests and response evaluations. In card systems, this is provided by a stub, which is detached and sent to the user's library. Holes punched in the stub can be read by computer, which can then prepare document order forms and tabulate the user response data.

The advantages of both listings and cards can be combined. A computer-printed listing of selected references can be accompanied by a stack of electronic data processing cards, which have been prepunched and interpreted with the document and user's identification. The user selects the cards that correspond to the announcements he has just read, punches out the appropriate prescored holes to express his interest evaluation and to request copies of desired documents. Returned to his library, the cards can serve as links in a fully automated system.

5. USER FEEDBACK

Optimum SDI service depends primarily on the user's interest profile and its improvement through feedback. It is important to understand the meaning of *optimum* service. Clearly, you as a user would best be served if, of the announcements you receive, all refer to documents that are definitely of interest to you. The announcements you receive should include every one in the file that would be of interest to you if you had a chance to

review it. Unfortunately, these are mutually incompatible goals. If you attempt to express your interests by structuring your interest profile in rather broad terms, some documents of no interest will be announced to you because of the various meanings that the indexer might have attributed to these terms while indexing the documents. If you attempt to be very precise in your choice of profile terms and further restrict their selective power by requiring Boolean intersections between terms, then you will miss being informed of some documents that you might have found of interest. Information scientists speak of the "relevance ratio" of (1) the number of documents of interest divided by (2) the number of documents that are announced, and the "recall or coverage ratio" of (1) number of relevant documents announced divided by (2) the total number of relevant documents in the system. In a very good system, you might find that 75 per cent of the announcements you receive are of interest, while these are perhaps 90 per cent of the relevant documents that are in the input data file.

Fortunately, the relevance and recall ratios can both be raised, although never to 100 per cent, by careful attention to the interest profile. By tabulating the responses that the user has fed back into the system, the operator can determine the profiles that need improvement. Successive tabulations and responses to user questionnaires reveal the success of the improvement effort. Furthermore, some users might be satisfied with one of the extremes -- either a broad announcement service giving all the documents the user can absorb, or a narrow selection of particularly interesting documents. The effort required to optimize the interest profile is the price paid for not having to look at every single announcement in an abstract journal with thousands of entries.

6. TREND TOWARD STANDARD PROFILES

When one examines the SDI systems mentioned so far, it is obvious that they possess certain features that are undesirable in the framework of providing information service to very large numbers of users. For one thing, each new user enrolled in the system adds to the requirements for computer time. Depending on the computer program, this increase need not be linear with number of users; nevertheless, the added cost and availability of computer time must be considered in planning any SDI system expansion. User turnover can be high in an SDI system and updating of the user profiles is a constant activity, again adding to computer usage. Besides computer costs, professional assistance in structuring interest profiles increases with number of users. The effort may well be justified in relation to the value of the SDI service, but the availability of professional personnel may be a problem.

One solution is the "group profile". Identical in every other respect to the individual interest profile, it selects announcements for an organizational unit; e.g., a branch or section. The unit has the responsibility of circulating the announcements so that all its members can select documents they wish to see. The group profile avoids duplication of interests between individual profiles, is not affected by personnel turnover, and because of its relative stability can be improved to an optimum level of performance more readily than can the number of individual profiles it might replace.

A second evolutionary development arising from SDI is the trend toward standard topical profiles. As with group profiles, the SDI match and print programs are continued, but instead of tailoring a profile to a particular individual's interests, a series of profiles is written to select announcements according to certain topics of defined scope. The computer-printed output for these topics is reproduced by conventional printing processes, and the user received copies of the particular topic listings that, together, best provide announcements meeting his specific interests.

The rationale behind this trend to topic profiles becomes clear when we examine a collection of individuals' SDI profiles. We find that many users have fairly clear interests in relatively well demarked subjects; e.g., aerodynamics, supersonic transports, geomagnetism, welding, etc. These subjects can then be considered as topics for which profiles might be

established. Other common interests can be determined by comparison of SDI profiles in a type of factor analysis. The resulting clusters of terms representing interests common to a number of users can also be considered as topic profiles. Although there is no real necessity for deciding on a simple title for such clusters, in practice a short subject title is chosen, which is then limited as to coverage by a scope note.

Selection of topics can be based on experience with bibliographic requests by potential users of the current awareness service, and of course by consideration of the subject content of the input documents.

7. NASA/SCAN PROGRAM

Typical of this new type of SDI service is the NASA/SCAN Program (Figs. 6, 7). SCAN is an acronym for Selected Current Aerospace Notices. SCAN is a developmental program with limited participation at present, but it offers the possibility of providing a selective current awareness service, not to the few hundreds of individuals typical of an SDI system, but to tens of thousands of aerospace scientists and engineers. This is possible because SCAN is much less expensive than SDI as the consequence of transferring much of the over-all effort from the computer operations and profile refinement activities to the traditional and relatively inexpensive operations of printing and sorting.

SCAN inherits the great flexibility of SDI in possessing the capability of modifying the scope of topics through profile changes and of adding or deleting topics at will. However, this flexibility cannot be used arbitrarily in a system striving for both economy and user satisfaction. Choosing the catalog of topics to offer potential users requires a tradeoff between a number of factors: (1) computer usage, which increases with the number of topics; (2) reproduction and sorting effort, which increases with the number of topics and number of users; (3) user satisfaction, which increases with increasing number of topics as the user's interests can then be correlated more closely with a limited number of topics. A decision on a particular topic thus includes consideration of the number of users having common interests, the extent to which users' specific interests can be met by a finite number of topics, the number of announcements we desire to set as a minimum for a topic per issue output, and the maximum number of announcements we will accept for a topic output. Too many announcements force the user to spend an excessive amount of time reviewing his lists of notifications, the solution being to split the topic into more specific coverage.

As an illustration of flexibility of the present NASA SCAN service, topics include *Supersonic Transports*, *Clear Air Turbulence*, and *Aircraft Noise and Sonic Boom*. The latter two topics provide the user who has these very specific interests with only the announcements he wishes to see, while the *Supersonic Transports* topic provides a much broader range of coverage. This flexibility extends through the SCAN topics, which can overlap in coverage and can announce the same document under a number of appropriate headings, permitting the user to match his specific or broad interests by a minimum of notification listings.

The notification listings are prepared by offset reproduction of the master computer printout and are then sorted by the requirements for numbers of copies of each topic as submitted by the participating organizations for their individual users. Thus, the sorting effort is distributed, with the local participating organization having the responsibility for maintaining user records and sorting and distributing the incoming SCAN notification listings.

User participation in SCAN optimization is important, but is not accomplished in the same way as in SDI. The user need not mark every announcement as to his interest, there being no provision for constant feedback as in SDI. Brief questionnaires as to the relevance of announcements and solicitation of comments on the desirability of creating new topics or combining several existing ones, or splitting one with too broad coverage, provide adequate feedback for optimizing the relatively stable SCAN profiles.

SCAN appears to be the path that current awareness services will take to provide service to very large numbers of users. Several U.S. Government agencies are testing programs much like SCAN. Projected costs for large scale SCAN programs appear to lie in the range of \$10 to \$20 per user per annum, again depending on the details of the service provided.

8. AVAILABILITY OF SDI PROGRAMS

Practical aspects of establishing a selective dissemination service include obtaining the computer program to accomplish the SDI match and printing. The program will vary with the computer to be used and with the type of announcements being designed. Organizations with the requisite programming staff and the computer testing capabilities may design and write their own program. The advantages are possible high efficiency and complete understanding, obtained from actual experience, of the full potentialities of the program. Programming an SDI system can be a very large effort, however, and the over-all program writing and testing can cover a long period of time.

If the SDI system design is tied to an already existing computer, a program for SDI service may already have been written and be available from the computer manufacturer or from associations of users of that particular computer. While use of an existing program may restrain the design of an SDI system to some extent, this may be less of a restraint in practice than it may seem at first thought. Furthermore, an existing program may possibly be modified more readily than writing an original program.

9. SDI INPUT

We commonly think of an SDI service as based on an organization's own document assessment and indexing activities. However, bibliographic data on computer tapes are increasingly available in certain subject areas. Certain organizations, the *Engineering Index* being an example, are in the early phases of such activities, with tapes on plastics and electronics being issued to a limited number of companies on a contract basis. Among professional societies, the American Chemical Society has advanced to the stage of offering a variety of index and bibliographic data on magnetic tape for general purchase. Commercial information science and technology firms also sell computer tapes containing bibliographic data covering various subject areas.

10. PURCHASE OF SDI SERVICE

As an alternative to in-house operation of an SDI program, the purchase of such a service may be considered. Societies and commercial firms that sell bibliographic data on tapes will, as an alternative, run the tapes on their own computers against the SDI profiles supplied by the customer. At least one U.S. firm offers selected references to a wide variety of the journal and patent literature based on a fee schedule. The customer may request announcements of all current documents published by a given author or a specified organization, or having certain keywords in the title, or that have cited a previous reference or author.

11. SELECTIVE DISSEMINATION OF DOCUMENTS (SDD)

An alternative or adjunct to the dissemination of bibliographic references is the selective distribution of the documents themselves directly to individual users. Large firms and professional societies are particularly interested in this means of bringing reported research to the attention of those individuals who can best make use of it and also in reducing the numbers of copies of documents that must be warehoused while waiting for requests. Documents can be matched to users by computer, using SDI-type profiles, or by a topic distribution like SCAN. Papers of conferences sponsored by a professional society or

internal reports created within an organization are particular candidates for SDD, perhaps in conjunction with an SDI announcement service for external reports. SDD might also be broadened to the distribution of all accessioned documents. In this case, distribution might be in the form of microfiche copies for economy. While such a complete SDD system has been proposed, its benefits over SDI announcements have not been demonstrated to justify its added cost.

12. FUTURE DEVELOPMENTS

Looking ahead to the next developments in selective dissemination, we can foresee increasing use of direct access to the computer made possible by time sharing and improved console and display devices. The coming generation of SDI users may, instead of receiving a listing of selected documents, merely sit down at the console of a computer interrogation station, possibly located at a considerable distance from the computer, and merely press a few buttons to identify himself and enter the code for his SDI announcements. The announcements selected since his last such request would be displayed on the cathode ray tube screen before him. By pressing another button, while a certain document is on the screen, he could instruct his library, through a terminal located there, to send him a copy of the document. NASA has underway a continuing study of the remote interrogation of large document data files known as RECON (for *remote console*). Incorporation of this SDI capability is to be tested in the near future.

On-line bibliographic interrogation of the computer offers exceptional advantages in rapid improvement of SDI profiles, as changes can be made while the output from the previous profile is being studied. The changes in selected announcements resulting from profile revisions can be called up from the data files immediately, making iterative testing highly effective in optimizing profiles in comparison to the present limitations caused by batching responses and delay in computer runs.

As current phases in SDI development progress, we may expect considerable clarification in the interplay between (1) SDI as presently constituted -- a batch process computer operation followed by a printout of all announcements, (2) SDI as it might become with the proliferation of on-line computer systems and (3) SCAN as the archetype of a system for distributing printed announcement lists to numerous users.

To look even further into the future, we must take account of the rapid advances being made in the capacities and speed of computers, the ability of optical readers to input full text instead of bibliographic data, the increasing capability of computers to organize raw data, and the potential developments in display and on-line dialogue between man and the computer. The science and technology of information is certain to advance also, so that automatic content analysis of documents will become possible to replace or supplement the intellectual indexing of today. Factors of document significance and relationship to users' interests will be far more complex than today. The SDI user in the future will not merely receive a listing of relevant documents but will be alerted, through optical display, to new information. The information might be a condensation or formatting, in graphical form when appropriate, of data received directly from experiments and that have not yet had a printed existence outside the computer. The printed document will still exist in abundance, but the SDI user will be alerted to the informational content rather than to the existence of the document itself. Furthermore, the portions of new information will be presented, by computer analysis of relation to the user's interests, in order of significance and immediacy of application.

While progress may appear to be slow at time, we are moving toward these potentialities, the ultimate goal being the enhancement of the users capabilities in advancing science and technology by a true communication of information.

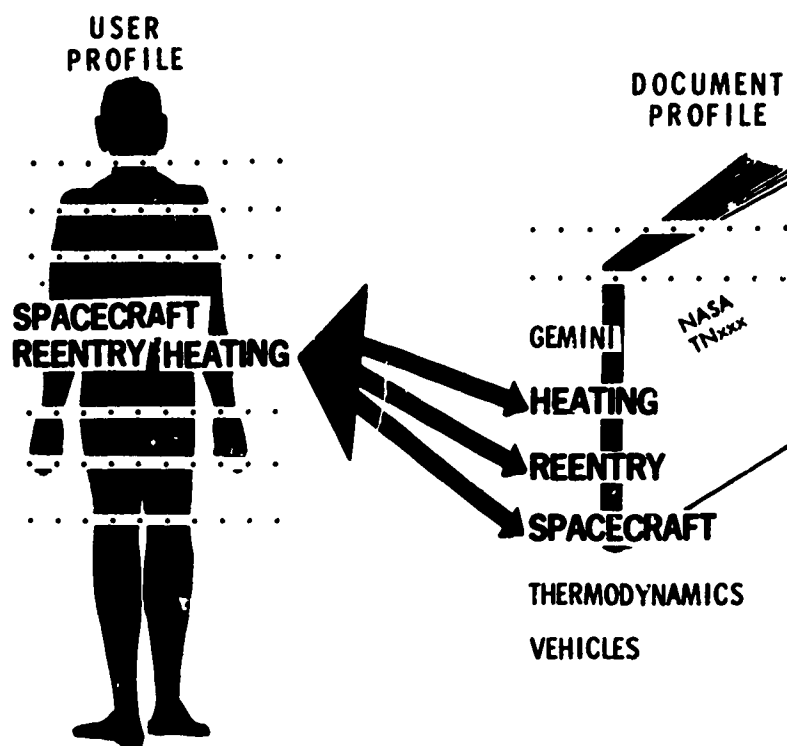


Fig.1 SDI: how it works

AM0101	*1	A E ANDERSON AME0101
	*2	N-227-1
	29	A001, C001
	32	CIT,NOC,TER
40 A	01	AERODYNAMIC DRAG
40 A	01	AERODYNAMIC STABILITY
40 A	01	FRICTION DRAG
40 A	01	HYPERSONIC AIRCRAFT
40 A	01	HYPERSONIC VEHICLES
40 A	01	LIFT
40 A	01	LIFT DRAG RATIO
40 A	01	LOW ASPECT RATIO WINGS
40 A	01	M WINGS
40 A	01	OGE WINGS
40 A	01	REENTRY VEHICLES
40 A	01	SKIN FRICTION
40 A	01	SLENDER BODIES
40 A	01	SLENDER WINGS
40 A	01	SONIC BOOMS
40 A	01	SUPERSONIC DRAG
40 A	01	SUPERSONIC TRANSPORTS
40 A	01	TRANSPORT AIRCRAFT
40 A	01	VORTEX BREAKDOWN
40 A	01	WAVE DRAG
40 A	01	WING CAMBER
40 A	01	WING PLANFORMS
40 C	-01	LIQUIDS
40 C	-01	METEOROLOGY
40 C	-01	SINKS
40 C	-01	THIN FILMS
40 C	01	VORTICES

Fig.2 SDI user interest profile


<p>1005-20007* # National Aeronautics and Space Administration Goddard Space Flight Center, Greenbelt, Md MAINNED SPACE FLIGHT NETWORK POSTMISSION RE- PORT FOR RANGER C AND D 26 Jul 1965 12 p (NASA-TM-X-55243, X-552-65-307) CFSTI HC \$100/MF SO SO CSCL 178</p> <p>Coverage of the orbital portions of the Ranger C and D missions through loss of signal (LOS) is analyzed. A brief critique of the performance of the Manned Space Flight Net- work (MSFN) for the two missions is presented with emphasis on troubles experienced. The general network requirements were to provide real-time computation support through in- section and LOS, to provide C-band radar beacon tracking through LOS, and to receive and record the Agena telemetry link until battery decay or retromaneuver. All of the MSFN stations that participated in the missions are listed. The network mission preparations are reviewed and the performance of the basic on-station systems and the computing and ground com- munications systems are summarized R N A</p>		<p>FROM NASA/SDI P. O. Box 5700 Bethesda, Maryland 20014</p>		<p></p>		<p>INSTRUCTIONS:</p> <ol style="list-style-type: none"> 1. Read the abstract 2. Respond by pushing out the appropriate box 3. Return this card to your library. 		<p>100 004794 FACILITY FORM 60-4</p>	
<p>Of Interest, Document Requested..... <input type="checkbox"/></p> <p>Of Interest, Document Not Wanted..... <input type="checkbox"/></p> <p>Of Interest Have Seen Before..... <input type="checkbox"/></p> <p>Of No Interest..... <input type="checkbox"/></p>		<p>Push Out This Box When Writing Address Changes or Comments Below..... <input type="checkbox"/></p>		<p>NAME JJ JONES MAIL STOP 199-10 XYZ</p>		<p>DEPT. LOCATION</p>		<p>ADDRESS CHANGE OR COMMENTS</p>	

Fig. 3 NASA two-card SDI announcement.

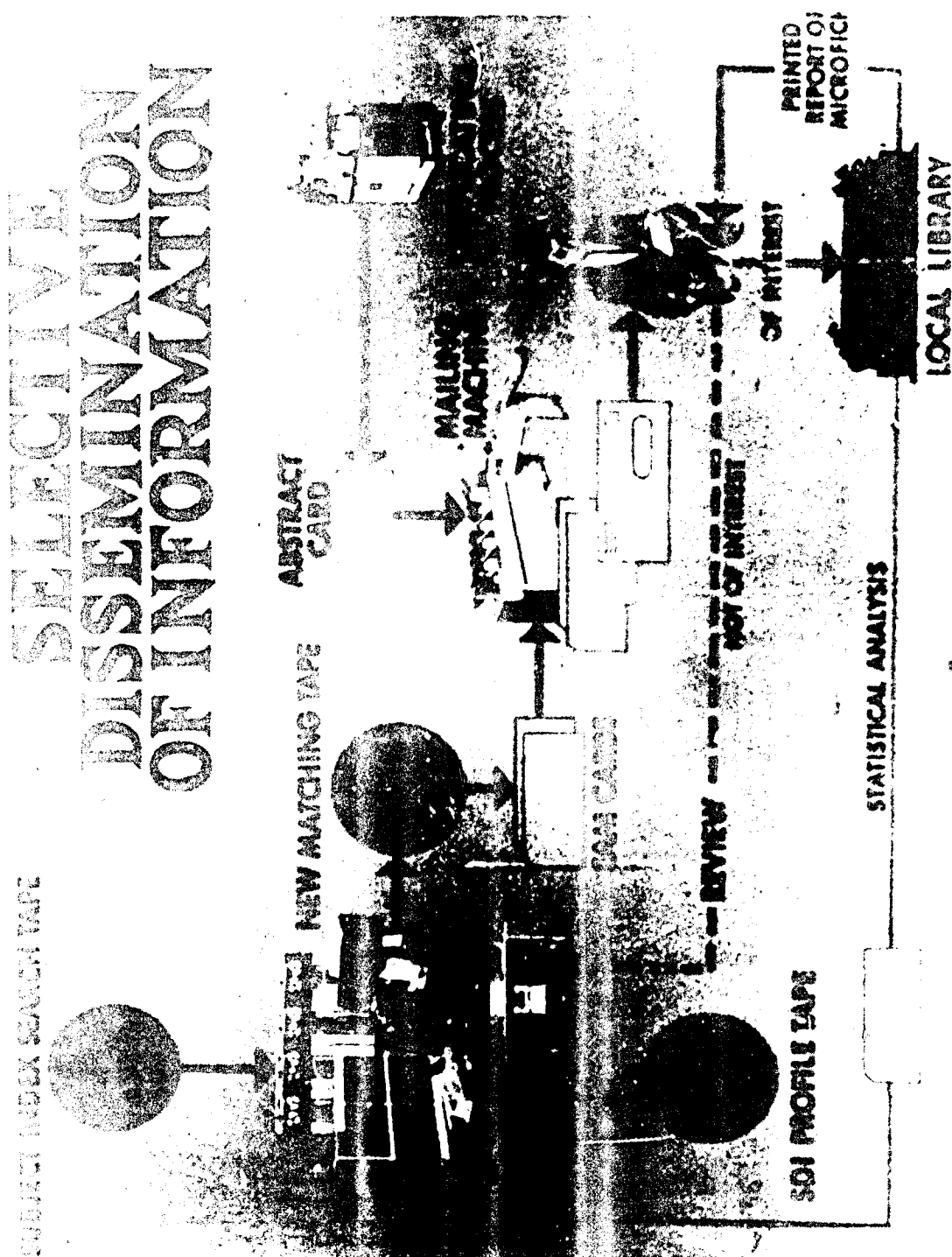


FIG. 4 Flow chart of two-card SDI system.

Best Available Copy

NASA / SDI NOTIFICATION

Please check the appropriate box. USER keep the top copy,
and send the second and third copies to your library.

NO INTEREST
OF INTEREST,
NOT REQUESTED
DOCUMENT
REQUESTED

STAR ISSUE #06, 23 MARCH 1968
A E ANDERSON AME0101
N-227-1

- N68-15019 #INSTITUT FRANCO-ALLEMAND DE RECHERCHES, ST. LOUIS /FRANCE/.
CAT. 02 BOOMS PRODUCED BY A MIRAGE 3 B IN ACCELERATED FLIGHT - OPERATION JERICO- FOCALIZATION IN ISTRES FROM 8-16 DECEMBER 1966
FLUGZEUGKNALLE EINER BESCHLEUNIGT FLIEGENDEN MIRAGE III B, OPERATION JERICO-FOCALISATION IN ISTRES VOM 8.-10. DEZEMBER 1966
FRANCESE, P. DATE- 10 AUG. 1967 COLL- 61 P REFS LANG- IN GERMAN
ISL-T-30/67
PEAK PRESSURE OF SONIC BOOMS PRODUCED BY MIRAGE 3 AIRCRAFT MEASURED OVER FIXED GROUND RANGE
*ACOUSTIC MEASUREMENTS, ELASTIC WAVES, FLIGHT PATHS, JET AIRCRAFT NOISE, MIRAGE 3 AIRCRAFT, *PRESSURE DISTRIBUTION, PRESSURE RECORDERS, *SONIC BOOMS, SUPERSONIC FLIGHT
- N68-15022 #CORNELL AERONAUTICAL LAB., INC., BUFFALO, N. Y.
CAT. 11 THE MULTI-RECOMPRESSION HEATER, A NEW CONCEPT FOR LARGE SCALE HYPERSONIC TESTING
WEATHERSTON, R. C. DATE- DEC. 1967 COLL- 61 P REFS
CAL-AD-2390-Z-1
THERMODYNAMICS, HEAT TRANSFER, AND MECHANICAL DESIGN OF MULTI-RECOMPRESSION HEATER FOR SIMULATION TESTING OF HYPERSONIC VEHICLES
*ATMOSPHERIC ENTRY SIMULATION, *CONVECTIVE HEAT TRANSFER, *EXPERIMENTAL DESIGN, *HEATING EQUIPMENT, HYPERSONIC VEHICLES, HYPERVELOCITY WIND TUNNELS, THERMAL ENERGY, THERMODYNAMICS
- N68-15099 SUD-AVIATION, PARIS /FRANCE/.
CAT. 02 CONCEPTION OF THE AIRBUS AND AERODYNAMIC PROBLEMS OF THE AEROBUS
CONCEPTION DE LA CELLULE ET PROBLEMS AERODYNAMIQUES DE L'AIRBUS
RCCHE, C. DATE- 1967 COLL- 23 P LANG- IN FRENCH CONF- PRESENTED AT A.F.I.T.A.E. 4TH COLLOQ. ON APPL. AERODYN., 8-10 NOV. 1967
EUROPEAN AEROBUS CONFIGURATIONS, PASSENGER TRAFFIC ECONOMICS, DEVELOPMENT COSTS, AND AIRPORT PLANNING
AIR CARGO, *AIRCRAFT CONFIGURATIONS, *AIRPLANE PRODUCTION COSTS, AIRPORT PLANNING, ECONOMICS, EUROPE, OPERATIONAL PROBLEMS, PASSENGERS, *TRANSPORT AIRCRAFT

Fig.5 NASA list-type, three-part SDI announcement form.

160002 *1 LASER APPLICATIONS
*2 SCAN 16-0002
29 A023, B001
32 ACC,CIT,TER
40 A 01 ANTENNAS
40 A 01 CAMERAS
40 A 01 COMMUNICATING
40 A 01 COMMUNICATION EQUIPMENT
40 A 01 COMMUNICATION THEORY
40 A 01 DOPPLER EFFECT
40 A 01 HIGH SPEED CAMERAS
40 A 01 IMAGE TUBES
40 A 22 LASERS
40 A 01 PHOTOGRAPHIC EQUIPMENT
40 A 01 PHOTOGRAPHY
40 A 01 RADAR
40 A 01 RANGE ERRORS
40 A 01 RANGE FINDERS
40 A 01 RANGEFINDING
40 A 01 RECEIVERS
40 A 01 SIGNAL RECEPTION
40 A 01 SIGNAL TRANSMISSION
40 A 01 SPACE COMMUNICATION
40 A 01 SPACEBORNE PHOTOGRAPHY
40 A 01 TELECOMMUNICATION
40 A 01 TRANSMITTERS
40 B 01 HOLOGRAPHY
40 B 01 LASER MODES
40 B 01 LASER OUTPUTS
40 B 01 OPTICAL COMMUNICATION
40 B 01 OPTICAL RADAR
40 B 01 WAVE FRONT RECONSTRUCTION

Fig.6 A standard NASA/SCAN topic profile.

NASA/SCAN Notification

03-04 HYDRAULIC AND PNEUMATIC SYSTEMS
TAA AND STAR ISSUES 905, MARCH 1968

Order the documents you want by checking the appropriate boxes. Then write your name and internal mail code in the spaces below, and forward the entire sheet to your library.

03-04

NAME

MAIL CODE

- 747 INCORPORATES PROVEN TECHNIQUES - NEW TECHNOLOGY.
PLATTNER, C. N. PUBL- AVIATION WEEK AND SPACE
TECHNOLOGY, VOL. 87, DATE- NOV. 20, 1967, COLL- P.
42, 63, 67, 69, 72-74.
AIR TRANSPORTATION, AIRCRAFT CONTROL, *AIRCRAFT DESIGN,
*CARGO AIRCRAFT, HYDRAULIC EQUIPMENT, *PASSENGERS, *WEIGHT
ANALYSIS, WING PROFILES C02 A68-15916
- A BISTABLE PNEUMATIC FLOW TRIGGER /PNEUMATYCZNY
STRUMIENIOWY PRZERUTNIK DNUSTABILNY/.
NICHALOWICZ, S. K. POLSKA AKADEMIA NAUK, INSTYTUT
AUTOMATYKI, WARSAW, POLAND PUBL- POMIARY,
AUTOMATYKA, KONTROLA, VOL. 13, DATE- DEC. 1967,
COLL- P. 552-555, 7 REFS. LANG- IN POLISH.
ACTUATORS, *BISTABLE CIRCUITS, *FLOW RESISTANCE, *PNEUMATIC
CIRCUITS, PNEUMATIC CONTROL, PNEUMATICS, PRESSURE EFFECTS,
*TRIGGER CIRCUITS C03 A68-16264 0
- THREE-DIMENSIONAL FREE JETS. RAJARATNAM, N.
SUBRAMANYA, K. ALBERTA, U., DEPT. OF CIVIL
ENGINEERING, EDMONTON, ALBERTA, CANADA PUBL- ROYAL
AERONAUTICAL SOCIETY, JOURNAL, VOL. 71, DATE- DEC.
1967, COLL- P. 858, 859, 7 REFS.
*FLOW VELOCITY, *FREE JETS, *HYDRAULICS, JET FLOW, LENGTH,
PREDICTIONS, *THREE DIMENSIONAL FLOW, VELOCITY DISTRIBUTION
C12 A68-16414
- DC-9 ENVIRONMENTAL CONTROL DESIGN AND FIRST YEAR'S
SERVICE EXPERIENCES. CLEEVES, V. F. MAUGER, H. M.
PERLEE, J. S. MCDONNELL DOUGLAS CORP., DOUGLAS
AIRCRAFT CO., AIRCRAFT DIV., LONG BEACH, CALIF.
CONF- /AMERICAN INST. OF AERONAUTICS AND ASTRONAUTICS,
COMMERCIAL AIRCRAFT DESIGN AND OPERATION MEETING, LOS
ANGELES, CALIF., JUN. 12-14, 1967./ PLAC- NEW YORK
PUBL- JOURNAL OF AIRCRAFT, VOL. 5, DATE- JAN.-FEB.
1968, COLL- P. 54-72. REAN- *FOR ABSTRACT SEE ISSUE
15, PAGE 2423, ACCESSION NO. A67-203740 AIAA PAPER
67-407
AIR CONDITIONING, *AIRCRAFT DESIGN, AIRCRAFT RELIABILITY,
CABIN ATMOSPHERES, CONFERENCES, *DC 9 AIRCRAFT,
*ENVIRONMENTAL CONTROL, PNEUMATIC CONTROL, PRESSURIZED
CABINS, *SYSTEMS ENGINEERING C03 A68-16602 0
- APPARATUS FOR SEMI-AUTOMATIC MEASUREMENT OF THE
LOGARITHMIC DECREMENT OF FREE VIBRATIONS OF GASTURBINE
BLADES. DZVICHENSKI, M. P. FASITZSKI, V. S.
TITOV, F. M. INIT- IN- INTERNAL FRICTION IN METALS
AND ALLOYS. EDITED BY V. S. POSTNIKOV, F. M. TAVADZE,
AND L. K. GORDIENKO. TRAN- /TRANSLATION OF VNIITRENNIE
TRENIYE V METALLAKH I SPLAVAKH, MOSCOW, IZDATEL'STVO
NAUKA, 1966./ PLAC- NEW YORK, PUBL- CONSULTANTS
BUREAU, DIV. OF PLENUM PUBLISHING CORP., DATE- 1967,
COLL- P. 211-216.
AUTOMATIC CONTROL, *FREE VIBRATION, GAS TURBINES, HYDRAULIC
EQUIPMENT, LOGARITHMS, *MEASURING INSTRUMENTS, *TURBINE
BLADES, *VIBRATION DAMPING C14 A68-16622
- HIGH ENERGY-RATE FORMING OF FIBROUS COMPOSITES.
ROBINSON, R. K. BATTTELLE MEMORIAL INST., PACIFIC
NORTHWEST LABS., RICHLAND, WASH. CONF- IN-
FIBER-STRENGTHENED METALLIC COMPOSITES, AMERICAN
SOCIETY FOR METALS, METALS CONGRESS, SYMPOSIUM,
CHICAGO, ILL., NOV. 2, 3, 1968, PAPERS. A68-16770
05-174 SPON- SYMPOSIUM SPONSORED BY THE AMERICAN
SOCIETY FOR TESTING AND MATERIALS, AND THE AMERICAN
SOCIETY OF MECHANICAL ENGINEERS. RESEARCH SPONSORED BY
THE BATTTELLE MEMORIAL INST. PLAC- PHILADELPHIA, PA.,
PUBL- AMERICAN SOCIETY FOR TESTING AND MATERIALS /ASTM
SPECIAL TECHNICAL PUBLICATION NO. 427/, DATE- 1967,
COLL- P. 107-123, 5 REFS.
ALUMINUM, ALUMINUM OXIDES, *COMPOSITE MATERIALS,
CONFERENCES, *FABRICATION, *FORMING TECHNIQUES, HONEYCOMB
STRUCTURES, PNEUMATIC EQUIPMENT, *REINFORCING FIBERS,
STAINLESS STEELS, TITANIUM, TURBINE BLADES C15 A68-16777
- RESEARCH AND DEVELOPMENT OF ON-BOARD SYSTEMS AND
ELEMENTS FOR AEROSPACE VEHICLES RESEARCH REPORT,
PERIOD ENDING 30 SEP. 1967 PENNSYLVANIA STATE
UNIV., UNIVERSITY PARK. SHEARER, J. L. DATE-
CCT. 1967 COLL- 13 P. REFS. NASA-CR-91679 RR-6
*AEROSPACE ENGINEERING, DIGITAL COMPUTERS, FLUID AMPLIFIERS,
FLUID JETS, FLUID TRANSMISSION LINES, *FLUIDICS, HYDRAULICS,

PAGE 1

LAST PAGE

Fig.7 NASA/SCAN notification listing.

DISCUSSION

C.O.Vernimb: Is the fact that NASA tries to do everything by computer the reason for changing from the tailored profile to the standard profile for SDI, with consequent lack of precision?

M.S.Day: The central organisation of NASA must provide an SDI service to 50,000 people and the present facility does not permit this to be done on the basis of tailored profiles so the idea of standard profiles has been developed. Local centres have accepted the idea for economic reasons.

R.J.Dubon: How does the sale of an SDI service stand with regard to the Copyright Laws?

M.S.Day: The NASA SDI service works only with titles and these are not considered to be copyright. If a system was set up which disseminated more information, e.g. abstracts, there might be the possibility of copyright infringements, but NASA's legal advisers do not think that infringements would in fact occur, even in this circumstance.

D.Bosman: In contrast to SDI services where users receive new information by post, the console-terminal system has the psychological disadvantage that positive action is required of each individual who participates in the system. Is this likely to reduce the effectiveness of the system?

M.S.Day: NASA has been testing the use of remote consoles over a period of eighteen months among its own technical staff. The only problem occurred when they were removed; workers wanted them back to get information which they could not obtain in other ways.

R.Moser: (i) Does NASA send out microfiche as well as hard copy?
(ii) Is there any standardisation between NASA and the Defense Documentation Center.

M.S.Day: (i) NASA does distribute microfiche and these are standard with microfiche of the Department of Defense and US Atomic Energy Commission. Microfiche provide the regular method of distributing input material of all types (except copyright material) to laboratories and agencies. The cost of each microfiche is about 10 cents.
(ii) About 30% of the STAR bulletin consists of material supplied by the Defense Documentation Center. A computer tape of this material is fed directly into the NASA system.

PAPER 14

INTERACTIVE INFORMATION PROCESSING,
RETRIEVAL, AND TRANSFER

by

Professor J.R.C. Licklider

Massachusetts Institute of Technology, U.S.A.

SUMMARY

Describes the present status and trends of man-computer-interactive information processing, retrieval and transfer made possible by multi-access computers. Some of the promises and problems of interaction are examined. The main activity in this field in the U.S.A. is the development of hardware-software systems and subsystems. Examples are drawn from three projects, MAC, TIP, and Intrex at the Massachusetts Institute of Technology.

INTERACTIVE INFORMATION PROCESSING, RETRIEVAL, AND TRANSFER

J.C.R. Licklider

1. INTRODUCTION

Face-to-face communication among people in small groups usually involves short sentences, frequent interruptions, as many questions as declarative statements, and as much meta-language as primary content. Each member of the group stimulates the others and is stimulated by them; the communication is - in the current terminology of on-line computing - "highly interactive". In contrast, a lecture tends to be one long message transmitted in one broad unidirectional channel; the poorer the lecture, the stronger that tendency. An ordinary document, consisting of passive print on passive paper, is not active and certainly not interactive at all. Moreover, the same is true of ordinary catalogues, indexes, abstracts, and accession lists and all the other traditional aids for finding documents - except librarians.

Librarians were joined recently by computers. The computers presented themselves at first as clerical assistants. They proved that they could help greatly in handling the routine chores of the library and the document room. Then they asked for and received raises and promotions, and they took on such additional work as making KWIC indexes and searching files of index terms. But that was just to get acquainted. What the computers really came to do is much more revolutionary: they came to do in a new and different way what only brains had done before - to make stored information interact.

Of course, "hardware" computers cannot do such a thing all by themselves: they must have "software", i.e., programs and data. They must also have input-output equipment through which they can interact with people. And they must have human users who know how to interact with them. Even when those requirements are met, present-day computers cannot (for lack of storage capacity) take the places of books and journals, but they can make available for interactive use such less voluminous information as the data of data banks and the citations contained in journal references.

Current experiments with programmed multi-access computers and computer-stored information are making it clear that interaction adds a very significant dimension to information processing, retrieval, and transfer. The purpose of this paper is to communicate some of the spirit and substance of that new dimension. The experiments and experiences described are from three research and development projects at the Massachusetts Institute of Technology: Project MAC, Project TIP, and Project INTREX. "MAC" stands for "Machine-Aided Cognition" and "Multi-Access Computers". "TIP" stands for "Technical Information Program". "INTREX" stands for "INformation Transfer EXperiments".

Research and development efforts in the field of interactive information processing, retrieval, and transfer are being carried out also in many other institutions in the United States as well as in Western Europe and the U.S.S.R. In selecting my examples from the projects at M.I.T., I am following a suggestion made by the organizing committee.

2. INTERACTIVE INFORMATION PROCESSING

By far the greatest part of the experience in interacting directly with computers - in interacting "on line" and "at the console", to use two of the favorite terms of the field -

comes not from experiments in information transfer but from use of computers in preparing computer programs and in solving scientific and engineering problems. The earliest digital computers were programmed on line, and there has always been a bit of on-line programming and "debugging" (elimination of program errors), but not until the advent of multi-access computing, based on the technique of "time sharing", was it possible for large numbers of people to work as a matter of course, day after day, at computer consoles. Now there are several hundred experienced console users at M.I.T. and many times that number elsewhere. It is something of an extrapolation to go from interactive programming and problem solving to interactive information retrieval and transfer, but the experience in the former is the main guide for experiments and developments in the latter.

At M.I.T. there are several interactive computer systems. The most widely used are two almost identical systems using IBM 7094 computers and a supervisory program called the "Compatible Time-Sharing System" (CTSS). Project MAC is operating a new system based on a Digital Equipment PDP-6 computer and is developing another new system that uses a General Electric 645 computer and a supervisory program called "MULTICS". There are two interactive systems in the Lincoln Laboratory, the TX-2 computer with a time-sharing supervisor called "APEX" and an IBM 360/67 computer with a time-sharing supervisor called "CP", and one in the Civil Engineering Department based on an IBM 360/40 with an auxiliary IBM 1130. However, the examples I shall use come from 7094-CTSS's^{1,2,3}.

Connected through a telephone switchboard to the 7094 computers are about 200 consoles. Most of the consoles are merely typewriters - i.e., teletypewriters or computer typewriters. A few have cathode-ray displays, typewriters, and auxiliary apparatus for graphical or manual-control input. A new console with storage-tube display and typewriter keyboard recently made its appearance⁴. It is rated high because the storage tube provides reasonably good resolution and fast display of text and/or drawings yet does not require continual "refreshing" to maintain the picture - and because it is less expensive than earlier consoles with cathode-ray displays.

The procedure for beginning a session at the computer is simple and has been described frequently elsewhere, so I shall assume that you (the user) have completed the formalities, have read the "mail" (messages) sent to you through the computer by your colleagues, and are ready to begin work. For the sake of simplicity, I shall assume that your console is a typewriter. Since CTSS is a general-purpose system with about a hundred different sets of programs ready to help you, what you do first depends upon the kind of work you have before you. The most trivial thing to do - and therefore a good introductory task - is to write a memorandum. What the computer will do for you is make it easy for you to correct typing mistakes and to effect editorial corrections, and then type out a "clean copy" of your memo. In order to prepare a memorandum, you call the writing and editing program TYPSET⁵ by typing "typset", and then you type the name you want to give the file you are going to create - for example, "jsmith".

The computer then types "W1720.2" (which means "Wait a moment." and "It is now 2/10 of a second past 17:20 o'clock.") and then "R000.2 + 000.1" (which means "Ready" and "You have thus far used 2/10 second of processor time and 1/10 second of drum-core transfer time.") and then "Input" (which means "I am ready for input from you.") You start to type the memorandum:

To: J.R.Smittyh
 .space
 Subject:
 @Subject: Plans for Improvement of planning Strategy
 .Space

Next Tuesday is the last day for###to submit your ideas

The character (#) ordinarily means to erase a character. The character (@) ordinarily means to erase a line. The control word ".space" means to skip a line. Your memorandum therefore stands as:

To: J.R.Smith

Subject: Plans for Improvement of planning Strategy

Next Tuesday is the last day to submit your ideas

You notice that you forgot to give the date and that you need to capitalize the "p" in "planning". To go into "edit mode", you press the carriage return key twice. The computer thereupon types "Edit". You are supposed, of course, to know the control words (or characters), only a few of which appear in this example. You type: "i Date: June 15, 1968" (the "i" meaning "insert"), press the carriage-return key, and then, for format control, type ".space". Then, realizing that you want to centre the date, you type "t" (for "go to the top of the file") and ".centre" (for "centre the next line"). In order to capitalize the "p" in "planning", you type "l plan" (for "locate the character string 'plan' ") and "c/plan/Plan" (which changes "planning" to "Planning") and press the carriage-return key twice to go back to the input mode and complete your memo.

And when you have completed the typing - and have corrected all your mistakes - you file the memo by typing "file jsmith". The computer types "W1735.1" and then "Roll.3 + 008.1". You type "runoff jsmith". The computer waits for you to put a fresh sheet of paper into the typewriter. You press the carriage-return key. The computer types the memo perfectly at 15 characters per second.

Such are the mechanics. They are very convenient if you want to prepare a long paper and don't type well.

2.1 Programming

Most of the people sitting at consoles at M.I.T. are preparing programs. Many of them use the programming language called "MAD" ("Michigan Algorithm Decoder"), but language preferences vary widely. Translators or interpreters for about 25 programming languages are available through CTSS. FORTRAN and LISP (List Programming Language) are popular. Many of the languages are special "problem-oriented" languages, e.g., STRESS and COGO (Coordinate Geometry) used in civil engineering and DYNAMO used in setting up simulations involving difference equations.

Typically a programmer types the program he is preparing with the aid of an editing program already in the system. When he has completed enough of the program to permit a test, he files the completed part, translates (compiles) it, and sets it to running. To compile a MAD program named SORTER, for example, the programmer types "MAD SORTER". When the computer reports that it has finished the compilation, the programmer types "LOADGO SORTER", and the program is off and running.

Of course, the program usually doesn't work properly the first time it is tested. The trick is - if you can't write a perfect program on the first trial - to have the computer help you find the errors. Its help is most direct if the program you write is highly interactive. When you run it, it types (or displays, or does something) to you at frequent intervals, and you respond to it. You can tell when, and therefore where (and often how) it goes wrong, and as soon as you see the flaw, you can call the editor, correct the flaw, recompile the program, and test it again. If the fault is obscure, however, you may need to call a "debugging aid". It is another prepared program. It has its own set of mnemonically coded instructions, but it adopts the vocabulary you have used in writing your program and (if it is a sophisticated aid) lets you search for flaws in the "source" programming language instead of the "object" code that actually is executed by the computer.

In any event, it is a joy to prepare, test, and correct programs on line. When you get used to interactive programming, you can't imagine how programmers could have persisted so

long in using the traditional, primitive ways - until you realize that they may have persisted so long because it took so long to perfect a program.

2.2 Programs

Much of the programming done in Project MAC has been done to facilitate programming. Some of it has been done to understand programming and to clarify the basic concepts of program, data, and file structures. Some of it has been done to explore and foster applications of interactive information processing, many of which involve users who are not, and need not be, familiar with programming. The result of the effort is a vast system of programs - over a million words of public programs and 20 to 30 million words of private programs - that are available all the time, day and night, except for about 6 hours a week that are devoted to maintenance. One of the main aims of Project MAC has been to see what kind of a community would grow up around a multi-access computer with such an accumulated software resource. Let me illustrate the range of the program resources with a few examples, and then let us consider the nature of the community.

One of the main service programs is MAP, a Mathematical Assistance Program (Kaplow, Strong, and Brackett⁶). It carries out mathematical operations for you - makes transformations, solves equations. It handles algebra, trigonometry, differential equations, Fourier and Laplace transforms, and so on. It plots graphs in linear or logarithmic coordinate systems, whichever you specify. You do not have to know all about it to use it: it asks you questions until it "understands" your problem.

A recently completed system of programs (Moses⁷) solves even quite complex problems in symbolic (indefinite, non-numerical) integration and does so about as well as, and much faster than, a good human integrator.

A system of programs called ADMINIS (Pool, Griffel, and McIntosh) facilitates the preparation, maintenance, and use of data bases. It is used heavily in work in the social sciences, where it is often necessary to work with large collections of fallible or fragmentary information. ADMINIS is designed to facilitate the transfer from ink-and-paper files to computer files without glossing over irregularities or losing track of distinctions.

TEACH (Weizenbaum, Fenichel, and Yochelson) teaches computer programming. It was used last Fall in one section of the most elementary programming course at M.I.T., and the experience suggests that the computer can do most, if not all, of the instructing required to let students make effective use of computers.

Cyrus Levinthal⁸ has developed programs that display structural diagrams of complex molecules. When the display apparatus⁹ rotates the diagrams, one sees them as though they were three-dimensional. With the aid of a light pen, one can move or twist or bend or stretch the structures. Given any configuration, the computer can calculate the electrical binding forces within the molecule. Levinthal and the computer work in partnership - he contributing the knowledge and the intuition, it contributing the calculating power and the ability to store large amounts of data accurately and display it in a meaningful pattern - to solve molecular-folding problems that neither could begin to solve alone.

OPS (amazingly, not an acronym) is a large system of programs for interactive, incremental simulation and modeling. It was developed by Greenberger, Jones, Morris, and Ness¹⁰. OPS comes about as close as any programming system yet developed to facilitating the basic process of thinking. First, it provides a language in which you can define objects or entities and specify their properties and the relations you think hold among them. Second, it lets you set into motion the situation thus described and make it unfold, displaying whatever aspects of its behavior you select. Third, it records the history and prepares whatever summaries you specify. And, fourth, it lets you intervene at any time, modify anything you like, and then cause the simulation either to continue or

to start again from the beginning. In short, it lets you organize your thoughts in a definite, moldable, dynamic medium; it reveals the implications of your static description by converting it into observable, moving behavior; and it lets you change your mind as often as you need to in order to explore the consequences of alternative assumptions and conditions.

2.3 The On-Line Community

The foregoing paragraphs described a few of the hundreds of programs - enough of them, I think, to convey an impression of what is meant by the phrase used earlier, "accumulated software resource". Those programs that have been recognized as the most generally useful have been described and catalogued more or less well and made available to all users. Other programs, less generally useful or less well recognized, reside in personal files, but they too can be used by anyone who tracks down their authors and gets permission. Thus, each user of CTSS can take advantage in his own work of pertinent efforts that his predecessors and his colleagues have made - instead of writing again a hundred-times rewritten program he can take on new worlds to conquer. A similar accumulative process is beginning to operate in the domain of data. A researcher formulates a theory, casts it into the form of a computer-program model, collects pertinent data, and tests the theory by applying the model to them. Someone else comes up with an alternative theory, programs it, links it to the first researcher's files (with permission, of course), and sets the two models into competition. Then others get interested, construct new models, modify old ones, collect more data. The theories change but the data base accumulates, and the accumulating data almost automatically make the new tests more comprehensive and less expensive than the old ones.

Through such computer-facilitated human interactions, a new kind of research community is arising at M.I.T. It is of course only in an early, formative stage, and its shape cannot yet be discerned clearly, but there is little doubt that something significant is happening. The computer system is being used for communication as well as for computing. People send messages to one another through the system. They learn about one another's programs before the programs are completed, sometimes even before the programming has begun. They plan together in order to maximize mutual value. They strive for generality and compatibility. Programs and data in the public files are beginning to be regarded as publications.

A rough measure of the interdependence of the members of the CTSS community is given by the ratio of L , the number of links from one person's files to the files of others, to F , the number of files one person has himself. The ratio L/F has risen from near 0 to about 1 in about three years. With better documentation and a more convenient file-linking scheme, it may go to 5 or even 10 during the lifetime of the new MULTICS system.

As the local on-line community, centred upon a single multi-access computer system, has emerged from concept into actuality, the idea of a broader, geographically distributed community has taken form in the minds of several people. This idea involves inter-connecting several multi-access computer systems and combining their communities of users into a supercommunity. There are evidences of interest in that thought and some incipient action based upon it. In my expectation, geographically distributed computers and information networks will come into being during the next decade. If they do, their impact upon the process of information transfer may be great.

2.4 Some Conclusions Based on Project MAC's Experience

During its five years of operation, Project MAC has explored more areas of the broad field of information processing than I can summarize here: theory of computation, theory of automata, programming language and systems, large files and data bases, architecture and organization of multi-access systems, graphic processing and display, modelling and simulation, console design and human factors in man-computer interaction, networks of central and satellite computers, and diverse applications. Let me, nevertheless, attempt

to state what I consider the main conclusions pertinent to information storage and retrieval:

- (a) Information is a dynamic, living thing, not properly to be confined (though we have long been forced to confine it thus) within the passive pages of a printed document.
- (b) As soon as information is freed from documental bounds and allowed to take on the form of process, the complexity (as distinguished from the mere amount) of knowledge makes itself evident. Everything one does in an active informational environment is "complexity limited".
- (c) Man-computer interaction is the most hopeful of the available approaches to the mastery of informational complexity.
- (d) Even with the help of on-line interaction, however, one man cannot master a very large domain of information. It will take cooperative on-line teamwork - in tight organizations or in loose communities, depending upon the natures of the undertakings - to achieve significant solutions + the "big" problems of science, technology, industries, cities, nations, and alliances.
- (e) The basic thing in the user's concept of an interactive information system is the "name space" of the filing (i.e., memory or storage) system.
- (f) Although the term "time sharing" has achieved wide currency, the sharing of processor time is not fundamentally important. Much more important are memory sharing and communication. Thus, the aim of multi-access design should not be to make each user think he has a computer all to himself; it should be to immerse each user in a cooperative, interactive, computer-based community.
- (g) The importance of controlled access to files can hardly be overstated. Control has two aspects: facilitation of authorized access and protection against unauthorized access. Good fundamental design can foster both, but little can be done to ameliorate a basically poor filing system.
- (h) The importance of fast interaction makes itself felt when a problem gets complex. One can wait an hour for a response - or even a day or a week - if only a few packets of information are involved in solving the problem, but waits of even a few seconds are prohibitive if thousands of factors have to be assessed and fitted into a pattern before a hypothesis can be tested.
- (i) In a community with many interests, the "general-purposeness" of the general-purpose multi-access computer system has real meaning and significance. The system must lend itself to a great variety of applications and serve as ready host to diverse subsystems. Generality and open-endedness cost something, of course, but Project MAC's experience indicates they are well worth it.
- (j) Reliable operation is vital, and - since the reliability will not be perfect - effective "back-up" arrangements and recovery procedures are vital, also. Before they will invest their main intellectual capital in, or entrust it to, a multi-access computer system, people have to be confident that the system will be available when they want it and that it will not lose their valuable programs and data.

3. INTERACTIVE INFORMATION RETRIEVAL

Project TIP¹¹ uses the facilities of the 7094-CTSS multi-access systems. The main TIP data base is a growing collection of bibliographic data, presently from almost 100,000 journal papers in the field of physics. The TIP programs are programs for processing the data in ways formulated by the user during his interaction with the system: simple searches for titles containing specified terms, for example, or complex explorations

involving sorting, merging, comparison of citations, progressive definition of retrieval specifications, and so on. The programs now in use (occasionally by many people throughout the Institute and intensively by about 20 devotees at M.I.T. and elsewhere) are the third revision of a system developed over about six years. They are applicable to diverse files and have been used on personal and fiscal as well as on bibliographic files. NEWTIP, now in a late stage of development, will extend the domain of the searching and processing tools to a still wider class of data bases.

Because TIP is so pertinent to the interest of this Symposium, I shall give a couple of examples from typical TIP sessions at the console. Using TIP, you type in lower case and the computer types back to you in all capitals. To explain what some of the abbreviations mean, I shall insert comments in parentheses. First, from the TIP User's Manual¹², a very simple example:

```
tip (You type "tip" to evoke the TIP program.)
W1019.5 (Wait. It is 10:19 and a half.)
TYPE YOUR REQUESTS.
search annals of physics v.26 to v.28
find title pion not author boyling j.b. (Find all articles with titles containing
"pion" except those by J.B.Boyling, whose work you already know.)
output print title a i and l (One-letter abbreviations are adequate for TIP words, such
as "author", "identification", and "location". You could just as well have typed
"o p t a i l" to instruct TIP to type as output the specified information about the
items found.)
```

```
go (Go to work, TIP.)
```

```
ANNALS OF PHYSICS
VOLUME 26
VOLUME 27
J384 V027 P0079
DEUTERON PHOTODISINTEGRATION AND N-P CAPTURE BELOW PION
PRODUCTION THRESHOLD
PARTOVI F.
CAMBRIDGE, MASSACHUSETTS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LABORATORY FOR NUCLEAR SCIENCE AND PHYSICS DEPARTMENT
```

```
VOLUME 28
J384 V028 P0034
ANALYSIS OF THE PHOTOPRODUCTION OF POSITIVE PIONS
HOHLER G.
SCHMIDT W.
GERMANY
TECHNISCHE HOCHSCHULE KARLSRUHE
INSTITUT THEORETISCHE KERNPHYSIK
```

(No article meeting the specification was found in volume 26. One was found in volume 27. One was found in volume 28. "J384" stands for "ANNALS OF PHYSICS". If you ask for "pion" you will find "pions", also - but not if you ask for "pion*". Several other devices similar to that use of the asterisk may be employed in specifying the type of match.)

Now, suppose you know two articles on a subject in which you are interested. They are in *Physical Review*, volume 135, and they begin on pages 247 and 582. You want to see what else there is in that volume that is closely related. You use Kessler's technique of "bibliographic coupling"^{12,13}:

```
search phyrev 135
f share b phyrev 135 247 (Find articles that share at least one bibliographic citation
with the article identified here...)
f share b phyrev 135 582 (...or with the article identified here. Putting the "f"s on
separate lines signifies "or".)
```

o p i t a linkage (Output by printing the identification, title, and authors of each shared citation and also the identifications of the article(s) with which the citation was shared. "Identification" means "journal, volume, and page".)

PHYSICAL REVIEW

VOLUME 135

J001 V135 P0247

ELECTRON SPIN DOUBLE RESONANCE STUDIES OF F CENTERS IN KCL. I

MORAN P. R.

SHARED LINKAGE TO PHYREV V135 P0247

J001 V070 P0460	J001 V091 P1071	J001 V098 P1787
J001 V102 P0151	J001 V110 P0630	J001 V114 P1245
J001 V115 P1506	J001 V118 P1024	J001 V124 P0442
J011 V022 P0989	J011 V026 P1124	J011 V029 P1692
J030 V026 P0167	J031 V032 P0775	J046 V005 P0183
J052 V008 P0299		

J001 V135 P0316

DOUBLE-RESONANCE PHENOMENA IN THE GASEOUS LASER

CULSHAW W.

SHARED LINKAGE TO PHYREV V135 P0247

J001 V070 P0460 J030 V026 P0167

SHARED LINKAGE TO PHYREV V135 P0582

J001 V076 P0833 J001 V107 P1559 J096 V229 P1213

J001 V135 P0470

SPIN-LATTICE RELAXATION OF F CENTERS IN KCL. ISOLATED F CENTERS

FELDMAN D. W.

WARREN R. W.

CASTLE J. G., JR.

SHARED LINKAGE TO PHYREV V135 P0247

J001 V091 P1071

J001 V135 P0582

SPIN RELAXATION OF OPTICALLY PUMPED CESIUM

FRANZ F. A.

LUSCHER E.

SHARED LINKAGE TO PHYREV V135 P0582

J001 V076 P0833	J001 V098 P0478	J001 V105 P1487
J001 V107 P1559	J001 V108 P1453	J001 V115 P0850
J001 V123 P0544	J001 V132 P0712	J003 V067 P0853
J012 V036 P0135	J012 V037 P2504	J017 V031 P0986
J031 V034 P0589	J034 V176 P0045	J038 V011 P0255
J041 V001 P0052	J041 V001 P0054	J041 V005 P0373
J045 V047 P0460	J046 V003 P0009	J046 V003 P0372
J046 V008 P0009	J046 V008 P0529	J046 V009 P0011
J049 V007 P0277	J052 V049 P0127	J074 V028 P0646
J096 V229 P1213	J096 V241 P0865	J096 V246 P1522
J096 V254 P3829	J256 V004 P0177	J273 V006 P1148

J001 V135 P0591
STUDY OF SPIN-EXCHANGE COLLISIONS IN VAPORS OF RB85, RB87, AND
CS133 BY PARAMAGNETIC RESONANCE

MOOS H. WARREN

SANDS RICHARD H.

SHARED LINKAGE TO PHYREV V135 P0247

J001 V070 P0460

SHARED LINKAGE TO PHYREV V135 P0582

J041 V001 P0054

J001 V135 P0727

LINE SHAPES OF PARAMAGNETIC RESONANCES OF CHROMIUM IN RUBY

GRANT W. J. C.

STRANDBERG M. W. P.

SHARED LINKAGE TO PHYREV V135 P0247

J001 V091 P1071

J001 V135 P1046

FAST-PASSAGE EFFECTS IN THE NUCLEAR MAGNETIC RESONANCE OF FE57
IN PURE IRON METAL

COWAN DAVID L.

ANDERSON L. WILMER

SHARED LINKAGE TO PHYREV V135 P0247

J001 V091 P1071

J001 V135 P1068

SPIN-LATTICE RELAXATION IN FREE-RADICAL COMPLEXES

KRISHNAJI

MISRA B. N.

SHARED LINKAGE TO PHYREV V135 P0247

J001 V070 P0460

J001 V135 P1099

LOW-FIELD RELAXATION AND THE STUDY OF ULTRASLOW ATOMIC MOTIONS
BY MAGNETIC RESONANCE

SLICHTER CHARLES P.

AILION DAVID

SHARED LINKAGE TO PHYREV V135 P0247

J001 V098 P1787

J001 V135 P1498

FORCED TWO-LEVEL OSCILLATOR

SENITZKY I. R.

SHARED LINKAGE TO PHYREV V135 P0247

J001 V070 P0460

J001 V135 P1622

LATTICE SUM EVALUATIONS OF RUBY SPECTRAL PARAMETERS

ARTMAN J. O.

MURPHY JOHN C.

SHARED LINKAGE TO PHYREV V135 P0582

J046 V008 P0529

As the bibliographic data are being typed, you note that two of the coupled articles shared many citations with one, but none with the other, of the source articles. You note, indeed, that there were very few doubly shared citations. Picking out the two articles that did share citations with both source articles, you call the library to see whether or not they are available. As you wait for the connection, you decide to convert the journal-volume-page identification into author-and-title citations, and, as a first step, you isolate and save the shared items by typing:

```
f share b j1 135 247 and share b j1 135 582
```

(Putting both on the same line would have set up the "and" relation even if you had not specified "and" explicitly.)

```
o save all (As the output action, save the data in a personal file.)
```

```
name save file resona biblio (Name the personal file "resona biblio".)
```

From that point you would ask TIP to print out the authors and titles of the articles in resona biblio, and then you would go on to explore other ideas.

Even though the foregoing examples exercised only a few of the TIP commands, they probably provided enough "scenario" to convey a notion of how one works with TIP. The TIP commands may be combined in many different patterns. Users develop ingenious strategies for filtering out irrelevant articles without losing the ones they want. Having found a good search pattern, however complex, one simply names it and applies it thereafter by merely typing the name. Or, having found a particularly rich collection of references, he preserves them in a personal file for future use. If there is too much output for the typewriter to handle, he stores the output in a personal file and requests that it be printed on a fast off-line printer and delivered by messenger or mail. If the TIP data base does not contain all the material he wants to use, there is a way for him to create TIP-processible files of his own.

To understand TIP as an experiment, one must shift from the user's to the designer's point of view. The designer can create retrieval tools to the limit of his imagination, but he must make empirical tests to find out what works and what does not. The tests have to involve "real" users. It is essential to record and analyze what the real users do.

While the TIP programs that we have described make searches and print lists, other TIP programs take notes. They record the identity of each user and, in chronological order, the name of each TIP command he issues and a summary statement of its result. At any time, a TIP user can type a complaint or a praiseful comment and be sure that what he says is recorded in the computer store for Dr. Kessler's benefit. The data thus collected are periodically analyzed, and modifications and adjustments are continually made. Indeed, TIP has developed through a process of guided evolution, and NEWTIP can perhaps be regarded as a guided mutation.

While TIP's users sit at consoles under programmed experimental scrutiny, they do substantial work. A book⁽¹⁴⁾ and two review articles^(15, 16) have been based on TIP literature searches, and TIP is being used in several studies of the flow and dynamics of the scientific literature. The TIP programs were used to prepare and print a catalogue of the books in the Student Centre Library, and the programs are being used to collect, process, and prepare for publication a catalogue of the journal and periodical holdings of the M.I.T. libraries. Informal studies were recently undertaken with the American Institute of Physics and the National Library of Medicine to explore problems of operational as distinguished from experimental application.

Thus, at least some of the value of on-line interaction in information retrieval is being exploited. But the real value is yet to come. Most of the work done thus far with TIP has been severely hampered by the slow pace of typewritten output, even when the typewriter is driven at its top speed by the computer. (One of Brown's¹⁵ searches, for which the computer required a scant 4 minutes of processing time, took 2 hours and 20 minutes of output typing.) Now we are looking forward eagerly to cathode-ray displays. They

will make it possible for the first time to explore fully the possibilities of interactive information retrieval.

4. INTERACTIVE INFORMATION TRANSFER

Whereas Projects MAC and TIP are old enough to be working on second-generation systems, Project INTREX is too young to have completed its first. In discussing interactive information transfer, therefore, I shall have to describe aims and aspirations rather than completed experiments or current experiences^(16, 17).

The purpose of Project INTREX is to conduct experiments that will clarify design objectives, methods, and techniques for information-transfer systems of about 1975. Emphasis is placed on the word "experiments". The project is working on a "system", it is true, but the system is being created to support the experiments. Experiments have been planned⁽¹⁶⁾ in four main areas:

- (a) bibliographic access,
- (b) physical access,
- (c) fact retrieval, and
- (d) network integration.

Thus far the project has concentrated on the first and second areas, progress in which I shall describe briefly. When work is undertaken in the third area, it will deal with computer-program methods of deriving, from formal data representations and/or from natural-language text, definite answers to specific questions. In the fourth area, the aim will be to interrelate M.I.T.'s computer-facilitated information-transfer system with systems in other universities, in government, and in industry. But obviously the third and fourth areas will not become critical until work in the first and second areas has come near to fruition.

4.1 Bibliographic Access

The purpose of the part of an information-transfer system that deals with "bibliographic access", of course, is to take the user from the stage in which he has only a nebulous idea of what he wants to the stage in which he holds in his hand the accession numbers (or equivalent identifiers) of the documents that will satisfy his (now more sharply defined) informational requirements. Most of the INTREX effort towards that end is centred upon the concept of the computer-based "augmented catalogue". It is computer-based in that it resides within the store of a multi-access computer and is interrogated from consoles. It is augmented in that it contains much more information about each document than does the traditional card catalogue, and also in that it deals with journal articles, theses, and reports as well as books.

Members of Project INTREX are preparing a computer-processible catalogue (data base) that will contain about 50 "fields" of information about each of approximately 10,000 documents in materials science and engineering. The 50 fields include all the conventional bibliographic data, such as author(s), title, affiliation(s), abstract, and key words or descriptors. Beyond those conventional data, the 50 include such things as a description of the intended audience, an estimate of the level of difficulty, and "feedback" comments submitted by knowledgeable users. The reason for including so many kinds of information is not that anyone is sure they will all be helpful; it is to determine which ones actually are helpful enough to warrant inclusion in a future operational system.

The 10,000 documents are being carefully selected to cover areas in which research is especially active at M.I.T. and in which there are researchers who will contribute to the planned experiments. Much of the selection is being done by the research people themselves.

The first experiments will be made with one of the 7094-CTSS multi-access computer systems, but with special consoles designed, and now being constructed, in the Electronic Systems Laboratory, where much of the INTREX research and development is being carried out. The initial system of catalogue-processing programs has been completed, and a more advanced system is being prepared for use with the consoles, perhaps as early as this coming summer. Consoles will be located in the Materials Science and Engineering Centre and the Engineering Library, and the experiments will be conducted within the context of actual use.

4.2 Physical Access

Bibliographic access must of course lead directly to physical access, to actual possession of the required substantive (as distinguished from bibliographic) information. Ideally, the substantive information would be stored in and delivered through the computer system that handled the bibliographic information - and would be available to the computer's processor for analysis and transformation. Limitations of the present technology, however, make digital storage and processing of a library-sized corpus uneconomic. The course being followed by Project INTREX is therefore to hold the substantive documents themselves in a non-digital microform storage system associated with the computer, and to use the computer - in which, we now assume, the bibliographic identifications of the required documents have already been specified - to pick out the identified documents and to execute their delivery to the user.

Accordingly, images of the pages of the 10,000 documents are being made in microfiche, and a computer-controlled subsystem for picking out and scanning selected pages is being constructed. The electrical signals derived by the scanner will be transmitted through coaxial cables to consoles in the Centre for Materials Science and Engineering and there restored to the form of a readable image, either "soft copy" (ephemeral) or "hard copy" (permanent). Equipment for several of the operations involved has been purchased or built and is ready for test. Equipment for the other operations is under procurement or development. The over-all design provides for rapid, guaranteed physical access to any document selected through the bibliographic-access system - and for delivery of that document directly to the location from which the retrieval operation was initiated.

Plans call for experimental investigation of such interrelated factors as the speed, the form, the resolution, and the cost of physical access. The experimental equipment will provide for fast delivery of sharp images, either hard or soft, but in some of the tests controlled delays will be introduced and the resolution of the reconstructed images will intentionally be degraded. By varying the parameters and making measurements of preference and performance under conditions of actual use, optimal engineering compromises will be approached and design objectives for operational systems will be formulated.

5. ADVANCED EXPERIMENTS IN A LIBRARY CONTEXT

Looking beyond the experiments with the 10,000-item collection in materials science and engineering, Project INTREX is conducting design studies that postulate a corpus of a million documents. At the same time, the M.I.T. Engineering Library is being reconstructed in such a way as to provide for simultaneous operations in conventional and computer-based modes. Card catalogues, book stacks, reading tables, microform equipment, and computer consoles will be brought together in an arrangement designed for advanced experiments in an operational library setting.

6. SYNTHESIS AND PROSPECT

Throughout MAC, TIP, and INTREX, and indeed throughout M.I.T., there is a feeling that a great and fundamental change is taking place in the way men relate to information. The change has not yet progressed very far. Its effects are not yet pronounced. Nevertheless,

one can see signs enough to tell that the change involves the rules by which the game is played.

The force behind the change is the computer, of course, but it is not the same computer we have known these last 20 years. It is not the lightning calculator, not the indefatigable clerk. It is the computer cast as the mouldable and retentive, yet dynamic medium - the medium within which one can create and preserve the most complex and subtle patterns and through which he can make those patterns operate (as programs) upon other patterns (data) derived from nature or the works of other men.

To almost every imaginative mind that has sensed the power of the computer as an interactive medium, it is obvious that it will change the very nature of libraries and information systems in the years to come. At the same time, it is clear that those years will be many - for many years of exploring and experimenting, many years of programming and debugging, and many years of developing and testing stand between us and the effective harnessing of the computer's power.

For the next few years, mainly because of the limitations of memory capacity that have been mentioned, we shall have to be satisfied - in library applications - with direct, on-line interaction with bibliographic information, followed by old-fashioned reading of substantive contents. I think that interactive bibliographic searches will prove effective, even in relation to their cost, in locating pertinent substantive information. On the other hand, I think they will fall far short of solving the basic problem of information transfer.

The basic problem arises, I believe, *after* a person has the required documents on his desk. It arises when he tries to transfer the substance of those documents across what West Churchman has called the "brain-desk barrier". That is when a person really needs the help of the computer.

To solve the fundamental problem it is necessary to make significant advances in the representation of knowledge and in the processing of languages, both natural and formal. It is necessary to convert substantive as well as bibliographic information into computer-processible form and to store it in, and interact with it through, sophisticatedly programmed computers. We shall doubtless not solve the fundamental problem in the near future, but at last we can work on it. In my opinion, it is the problem that deserves our best and greatest efforts.

ACKNOWLEDGEMENT

This paper was prepared under the aegis of Project MAC, a research and development activity supported by the Advanced Research Projects Agency of the U.S. Department of Defense through the Office of Naval Research under Contract Nonr-4102(01). In addition to some of the work of Project MAC, the paper describes some of the work of Project TIP and Project INTREX. Project TIP is supported by the National Science Foundation under Research Grant GN-589. Project INTREX is supported in part by a grant from the Carnegie Corporation, in part by Contract NSF-C472 from the National Science Foundation, and in part by the Council on Library Resources, Inc. It is a pleasure to express appreciation to the sponsors for their support of the work described herein and to the directors of the three projects and the director of the Electronic Systems Laboratory for their cooperation during the preparation of the paper: Professor R.M.Fano, Director of Project MAC; Dr M.M.Kessler, Director of Project TIP; Professor C.F.J.Overhage, Director of Project INTREX; and Professor J.F.Reintjes, Director of ESL.

REFERENCES

1. Fano, R.M.,
Corbato, F.J., *Time Sharing on Computers.* Scient. Am. Vol.23, 1966,
pp.128-140.
2. Christman, P.A., (Ed.) *The Compatible Time-Sharing System*, 2nd Edition, Cambridge,
Massachusetts, The M.I.T. Press, 1965.
3. Fano, R.M., *The Place of Time Sharing*, J. Engng Educ., April, 1968,
(scheduled publication).
4. Stotz, R.H.,
Cheek, T.B., *A Low-Cost Graphic Display for a Computer Time-Sharing
Console.* Proc. 8th Nat. Symp. Inform. Display, 1967, pp.
91-100.
5. Saltier, J.A., *Typset and Runoff, Memorandum Editor and Type-Out Commands.*
MAC-M-193-2, 1-15, January, 1965.
6. Kaplow, R.,
Strong, S.,
Brackett, J., *MAP: A System for On-Line Mathematical Analysis.* Massachu-
setts Institute of Technology, Cambridge, Massachusetts,
Rep. MAC-TR-24 Project MAC, 1966.
7. Moses, J., *Symbolic Integration.* Massachusetts Institute of Technology,
Cambridge, Massachusetts, Rep. MAC-TR-47 Project MAC, 1967
(Thesis)
8. Levinthal, C., *Molecular Model-Building by Computer.* Scient. Am., Vol.24,
1966, pp.42-52.
9. Ward, J.E.,
Stotz, R.H., *Operating Manual for the ESL Display Console.* Massachusetts
Institute of Technology, Cambridge, Massachusetts, Rep.
MAC-M-217, Project MAC, and Electronic Systems Laboratory,
Massachusetts Institute of Technology, Cambridge,
Massachusetts, Rep. ESL 9442-M-129 1965.
10. Greenberger, M.,
Jones, M.M.,
Morris, J.H., Jr.,
Ness, D.N. *On-Line Computation and Simulation: The OPS-3 System.*
Cambridge, Massachusetts. The M.I.T. Press, 1965.
11. Kessler, M.M., *The M.I.T. Technical Information Project.* Physics today,
Vol.18, No.3, March, 1965, pp.28-36.
12. Kessler, M.M., *An Experimental Study of Bibliographic Coupling Between
Technical Papers.* I.E.E.E. Trans. PTGIT, IT-9, pp.49-51.
13. Kessler, M.M., *Bibliographic Coupling Extended in Time: Ten Case Histories.*
Inf. Stor. Retr., Vol.1, 1963, 169-187.
14. Ashburn, E.V.,
Lengyel, B.A., *Bibliography of The Open Literature of Lasers.* J. Opt. Soc.
Am., Vol.57, 1967, pp.119-148.
15. Brown, S.C., *A Bibliographic Search by Computer.* Physics to-day, Vol.19,
1966, pp.59-64.
16. Overhage, C.F.J.,
Harman, R.J., *INTREX: Report of a Planning Conference on Information
Transfer Experiments.* Cambridge, Massachusetts, The M.I.T.
Press, 1965.
17. - *Semiannual Activity Report.* Massachusetts Institute of
Technology, Cambridge, Massachusetts, PR-4, Project INTREX,
15 September 1967.

DISCUSSION

D. Bosman: What is the intellectual level of the people who use the MAC system?

J. C. R. Licklider: The system is used a lot by professors, research students and graduates on the academic side and also by others such as deans and administrative staff who find it useful for record-keeping. Only a few undergraduates have access to the system.

S. Skoumal: What are the difficulties about MAC which make it unattractive for commercial exploitation by private operators?

J. C. R. Licklider: The present MAC system operates on out-of-date hardware which makes it inefficient, but a commercial firm could almost certainly develop an efficient system and sell its services.

The main problem is that it takes a very long time to develop the software for the system and during this time the efficiency of the hardware available will have improved very considerably so that one tends to be operating always on outdated hardware.

I. Gabelman: What is the status of the MULTICS system?

J. C. R. Licklider: MULTICS (Multiplexed Information and Computation System) operates on a GEC 645 computer with associated software. At present it operates too slowly and current work is aimed at speeding it up. The system has already been under development for about three years so that by the time it reaches the level of operation of the CTSS (Compatible Time-Sharing System) the hardware will be out-of-date.

T. Einsele: Is the MULTICS system designed for a special group of users?

J. C. R. Licklider: We are developing the protocol for deciding on a community of users. The system will be available to various small computers each of which will serve as the interface with the MULTICS system.

R. Stark: Is there an index of the programs used in Project MAC and can this be made available to those interested?

J. C. R. Licklider: There is a program manual which contains text of all the programs and this could be made available at the cost of copying but it would be very difficult to utilise as about twenty-five different programming languages have been used.

PAPER 15

MAN-MACHINE INTERFACE

by

Professor W.Händler

Erlanger University, Nuremberg, Germany.

SUMMARY

The problem of the man-machine interface is traced back to the time when the first computers were designed. In overcoming the problems of the interface the cathode-ray tube display is of prime importance. Using the display screen it is possible to transmit almost instantaneously to man alphanumeric text, black and white shading, scaled shading, coloured pictures and moving pictures. The mathematical theory of automata is opening up new ways of examining the problems in a formal manner. Solution to many of the problems at the interface, however, still awaits better knowledge of how information processing takes place within the human nervous system.

MAN-MACHINE-INTERFACE

W.Händler

1. INTRODUCTION

The subject "Man-Machine-Interface" is a very broad one and some of its aspects cannot be regarded here. Much has been written on the more linguistic or symbolization point of view elsewhere. I shall restrict myself to some general questions about the interface which seem to be important in my opinion.

There is a very interesting novel by Samuel Butler written in 1872. The name of this novel is "Erewhon", which can be read almost inverted as "nowhere". The book tells of a traveller who discovers a strange country far behind some mountains. He becomes acquainted with the natives there and is surprised to find deserted railways and other evidence of decayed technological equipment. None of it is in operation. The original technological world seems to have died out. When people notice a watch belonging to the hero of our story they accuse him of violating their laws. The reason is that the people had a highly developed technology long before the arrival of our informant. One day the machines (we would say to-day: the automata) rebelled. They had become conscious of their existence and fought against men. Finally the people succeeded in defeating the machines by an extreme effort. From this day on all technology had been banned.

Let me mention only that our hero finally escaped alive.

I have told you this story only in order to make clear that the problems of relating man and machine seem to have come up previously in our century. We now realize that the interdependence of man and machine can be very close. Indeed, I fully agree with Dr. Licklider and other scientists of MIT who speak of an emerging man-computer community which may eventually embrace a vast network of different computers and human users.

However, the question as to whether the machine has a consciousness or tenders any friendly feelings towards us is fortunately fading out of the discussion. We are using the machine and especially computers in a strictly pragmatic way whereby we are possibly questioning our social and moral responsibility with respect to its use.

2. HISTORY OF MAN-COMPUTER RELATIONS

The first computers designed more than 20 years ago were very much characterized by the notion of entirely predetermined computational processes or algorithms set off by a single starting signal. Some years later the computer was equipped with additional switches, among them so-called selector-switches, for guiding the course of the program along alternative pathways.

At first the inventors themselves who were intimately tied to their creation operated the system. Soon other people came to use the machine and worked on the basis of a start-button philosophy.

Frequently there resulted a feeling of utter resignation in the face of the microsecond. The user could hope to trace computational processes only in a very crude, overall manner, perhaps merely with regard to the beginning and end of his job. However, note-worthy

efforts were made repeatedly in an attempt at re-establishing a somewhat closer contact with the ever faster growing computer.

With further increase in speed it became necessary to exclude direct user operation in favour of batch-processing thus saving costly computer time previously wasted on tedious input-output procedures. Trained people undertook the task of empirical scheduling. In this manner the problem of man-machine interface found a somewhat fictitious solution; the real user rarely came to see the computer or press any of its buttons. Man-machine-interface was restricted to a rather small group of trained personnel.

Consequently user and computer were alienated from each other, a small group of technicians - the operators - providing their sole common tie and most flexible but entirely human interface because man is still the most adjustable and adaptive instrument.

The disadvantage of the method described results from missing the rather vital experience of observing the machine undertake your assignment. If you are kept at a distance from the computer you may well take weeks to accomplish the objectives which you might achieve in minutes by direct contact with the machine.

This kind of disproportion can be corrected today by allotting a substantial number of consoles to various permanent customers in a time-sharing configuration.

3. PICTURE-REPRESENTATIONS OFFERED BY THE COMPUTER

Purely typewritten or teletyped communication is only one of the present possibilities and, having originated in a pre-computer age, not necessarily the most efficient one.

Prime importance perhaps should be attached to the development of displays of the cathode-ray-tube type which permit almost instantaneous transmission of mixed-mode text or picture information eliminating tedious waiting for carriage return or line feed motions of printer or teletypewriter. Via display-screen the computer is able to offer alternatives to be chosen effortlessly and swiftly with the aid of a lightpen, for instance. The same is true for the Rand-tablet.

In the past most transitions to advanced techniques have been marked by their initial tendency to simulate already existing achievements using the new methods. But when we employ the display consoles mentioned we will have to do some fundamental rethinking. Evidently the flow of information out of the computer can be speeded up well beyond the limits hitherto known. We must ask for the maximal amount of information per unit of time and the optimal manner of presentation best suited for processing by the human user.

There are several distinct categories of representation, such as alphanumeric text, black and white shading, scaled shading, coloured pictures, moving pictures, 3-dimensional displays. Those representations may occur separately or in different combinations with each other.

Determination of suitable quantities or qualities of information is exceedingly difficult. Many partially subjective factors are involved. Our judgement will also depend on the particular topic under consideration. Graphic representations may be of insignificant value when applied to codes of law or administrative regulations. Alphanumeric text on the other hand is not sufficient for describing processes of industrial production or technical design. If you interpret the list above as ordered according to quality it may be valid perhaps in a majority of cases but certainly not in general.

4. THE FACTOR "TIME"

Another important aspect must not be disregarded. Up to now most display applications have been of a rather static nature in spite of some rudimentary attempts at making

computer-controlled cine-films. Usually considerations have centred around the effects of single pictures rather than taking account of the complete sequence of textual and graphic information as a totality meant to establish dynamic contact between user and machine.

Lately a great deal has been said about lengthy waiting times leading to disturbed contact and distracting the user from his task. Conversely an excessive reaction speed of the computer may "overfeed" the user and cause in him what the psychologists call a "mental block".

The user tacitly assumes the computer expects reactions without delay and seems to feel pressed for time. Reasonable programming for time-sharing systems should avoid creating this feeling.

So far I have mentioned a number of psychological factors influencing man-machine interface. It seems to be clear that other more physiological factors also influence the effectiveness of man-machine-interrelations.

I remember the endeavour of my young colleagues to program the GOMOKU-Game or "Five in a Row" in the sense of an optimal CRT-dialogue between man and computer as the two partners of the game. After the computer has lost a game it retires for a certain period of time e.g. some seconds or even one minute in order to learn from the situation. But in all intervals requiring merely positioning a piece the computer does it within one millisecond.

The time the computer needs to calculate its decision is usually negligible when compared to human reaction speeds. In most cases man is not even able to observe what the computer does. Therefore he may not perceive which one of sometimes many pieces the computer has put on the board (i.e. the CRT-screen). In contrast the human partner will normally handle the lightpen only after a delay due to deliberation in order to place the next piece. So the human reaction requires some seconds or even some minutes. This interval would be important for an observing human opponent. It is generally useless to the computer.

Let us return to the computer projecting its pieces onto the screen at its own fast pace. In this case as in other cases I think we must correct the behaviour of the computer.

It is not always necessary to alter the computer's timing behaviour. Instead of this we can cause the computer to mark the last piece (Fig.1) by using a third sort of symbol ("last piece" of the computer). This provision substitutes the typical slow behaviour of a human partner.

Here is an example in which a symbolic representation must compensate human incapability to observe certain millisecond-events. Perhaps we should be able to formulate a law of interchanging time and space. In the example above space took the place of time by the introduction of this additional third type of symbol shown.

We see that in contrast to the man-man-interface there exists an entirely unsymmetric situation with respect to man-machine interface. We can assume in accordance with the theory of evolution that human beings naturally are best adapted to man-man-interface, which is highly symmetric.

5. THE UNBALANCED MAN-MACHINE-INTERFACE

Man-machine interface is unsymmetric both with respect to quantity and with respect to quality of the information processed. Man is able to deal with a highly complex visual supply of 10^9 bits per second (Fig.2, the diagram originated by Prof. Keidel, Universität Erlangen-Nürnberg). Human beings process this amount of information in a highly effective manner. As far as we know to-day there is a hierarchy in our nervous system filtering the information by a principle of reinforcement. This filtering process results in a residual flow of only 10^2 bits per second into our consciousness.

Preceding this region of consciousness we have at least three other levels in which the nervous system branches the information to motor systems in order to supply the "subroutines" there with parameters. Some of these subroutines can be simulated like the so-called conditioned reflex, an unskilled or built-in program. But other more complex programs - those more skilled - are not well understood yet. There are some hypotheses like the "perceptron" which are suited to shed some light on cognition or data reduction in the human organism.

Leaving the input to man and turning to his output, our investigations seem to show that man is able to produce information at the rate of nearly 10^7 bits per second including gesticular and other motorial transmission. These reactions are initiated within the region of consciousness at the rate of presumably 10^2 bits per second as described in Fig. 2. Then certain subroutines are triggered level by level and are supplied with parameters at the same time in the way indicated.

Man normally has only poor ability to manipulate figures exactly without recording those figures on a slip of paper. This process requires repeated outputs and thereafter inputs to an. It is in their adaptness for record-keeping and performing arithmetic manipulations that computers generally excel human capacity.

In contrast, man has the ability to evaluate intertwined complex structures qualitatively, sometimes perhaps in spite of the fact that he has never met with comparable phenomena before. It looks rather improbable that computers will ever acquire similar facilities for qualitative appraisal.

On the other hand the computer can offer an amount of coded or printed information, which can never be read by a single person. Also the computer is able to produce program-controlled moving pictures using its enormous calculating power for creating kinds of output man can never realise without computer aid. As an illustration of what I have in mind you may take the snapshot (Fig. 3) of a game played on a PDP-7 display unit. The game was invented at Cambridge University and simulates a sort of "naval battle" in which the opponents attempt to hit each other's "ships" (circular bright objects appearing to move in a viscous fluid) with "rockets" (sparks of light).

We have mentioned above that man may pour out up to 10^7 bits of information per second. This performance might be exceeded by the computer some day. The figure 10^7 however is not a very realistic value for man when applied to lingual and grammatical formulations in order to express effectively some of his thoughts. Therefore it seems to be more reasonable to assume the proper figure of man's output around 10^2 bits per second at most.

On the other hand no computer to-day is able to catch any T.V.-like pictures instantaneously as we do. Our experience with pattern recognition facilities is not of the sort we need for the serious use of such features by a computer aside from some special applications.

To-day, the computer still requires coded or at least prepared information within the limits of rather strongly restrictive conventions.

6. ARE THERE SUGGESTIONS FOR IMPROVING MAN-MACHINE-INTERFACE?

The subject of this meeting "Storage and Retrieval of Information" suggests that we try to compare man's ability in this field with the ability of the computer. We must acknowledge man's great superiority over the computer in the whole area with the exception perhaps of his weakness in memorizing long sequences of numbers. Scarcely do we get any suggestion today at all as to how we should really proceed with data reduction, storage and retrieval of information in a manner analogous to the methods employed by living organisms. We have not succeeded in finding the enormously effective principle of nature. At present - for example - there seems to be no possibility of localizing the content of human memory. The same applies to the processes of association and of generalization in the human brain.

These factors certainly influence man-machine-interface, but they are not part of man-machine-interface, in a certain sense.

Let me summarize the most important facts as I see them for improving man-machine-interface.

- (a) As in supervisory programs we should install open-ended modular blocks of prefabricated processes. These processes must be supplied with parameters automatically after having been started (similar to the process in the organism).
- (b) The predominant idea of an algorithm which has to be started and then has to run off without taking care of any signals and parameters from outside should disappear in favour of an adaptive conception in programming. The computer would be subject to an internal process of evolution.
- (c) Apart from other important questions such as the associative memory we should investigate eventually the possibilities of connecting nerve fibres of man with the computer directly in order to initiate complex subroutines in critical cases when prompt responses are required. Related important applications come up in medicine.

I should like to remark finally that the problems in my opinion are mainly not of a conventionally technological kind. Rather we must understand more completely the structure of organic behaviour and mental processes.

7. SOME REMARKS ABOUT AUTOMATA THEORY

The question arises whether there are tools suited for dealing with man-machine-interface in a more theoretical way. For instance the IVERSON-Language (A.P.L.) like other languages has the disadvantage of essentially disregarding the state of the automaton.

In my opinion a rigid and formal access to the subject is opened by the mathematical theory of automata. Here the user is considered to be a finite automaton in the sense of a working hypothesis. Finite automata are capable of occupying a finite number of states. On reception of an input symbol they pass in a well-defined manner (from the deterministic or stochastic point of view) from one state to another emitting an output symbol. Present-day theories however are still inadequate to deal successfully with the formidably large number of states actually existing in both the computer and the human organism.

Further evolution of Automata Theory will certainly improve upon this deficiency. There will be the possibility of treating very complex structures with the aid of algebraic concepts like homomorphisms, groups of automorphisms and so on.

The subject man-machine-interface is at present far from being formalized in a way familiar to me. But nevertheless some advantages of other applications of the Theory of Automata encourage us to make efforts in this direction.

Within the framework of present-day theory the automaton for instance is called upon to establish equivalence classes of patterns (e.g. the class of all syntactically correct sentences of a language) or - less ambitiously - to make decisions regarding the membership of a given pattern to one of those classes. Another part of the theory is devoted to the design of a suitable series of experiments for determining the initial state of an automaton. These investigations can and should be extended to the reciprocal effects of two automata on each other - their dialogue - and to the mutual interaction of many automata - their communication, or even their social behaviour.

The axioms and conclusions of mathematical theory will depend on the fundamental concepts underlying our comprehension of man-machine interface.

One view is dualistic. It treats the machine as man's conversational partner with special abilities for solving certain problems.

The other view can be called monolithic and teleologic in a way. Man and machine are thought of as a single integrated system organized for the purpose of solving a problem or mastering the environment.

The latter notion is unfortunately subject to severe practical limitations because - as I have already said - we still possess too little knowledge of how information processing takes place within the human nervous system.

Once deeper insights into these mechanisms have been gained we stand a good chance of designing very effective man-machine interfaces on the basis of the concept last mentioned.

DISCUSSION

S. Skoumal: Is the research in this area carried out only at universities or is some interest shown by industry?

W. Händler: At present very little work is done and then only in universities. The interest of industry would be welcomed.

P. Molzberger: Have any experiments been done coupling human nerve fibres with computers?

W. Händler: Although there has been much co-operation between the physiologist and the mathematician this experiment has not been attempted as far as I know.

J.P. Little: Are any investigations being made into the possibilities of feeding speech directly or indirectly into the computer.

W. Händler: I am not aware of any investigations. Almost certainly some sort of pre-processing of speech would be required before signals could be passed to the central computer.

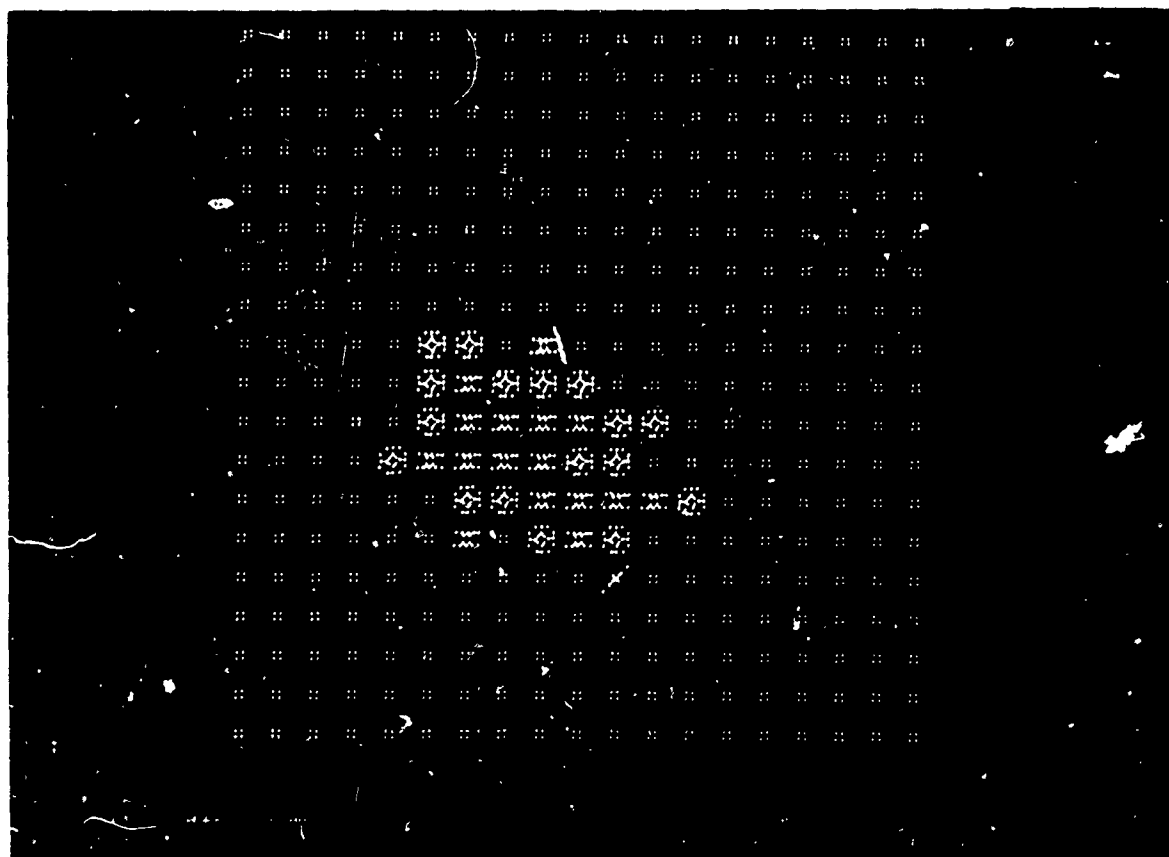


Fig.1 CRT-display of GOMOKU-Game or "Five in a Row"

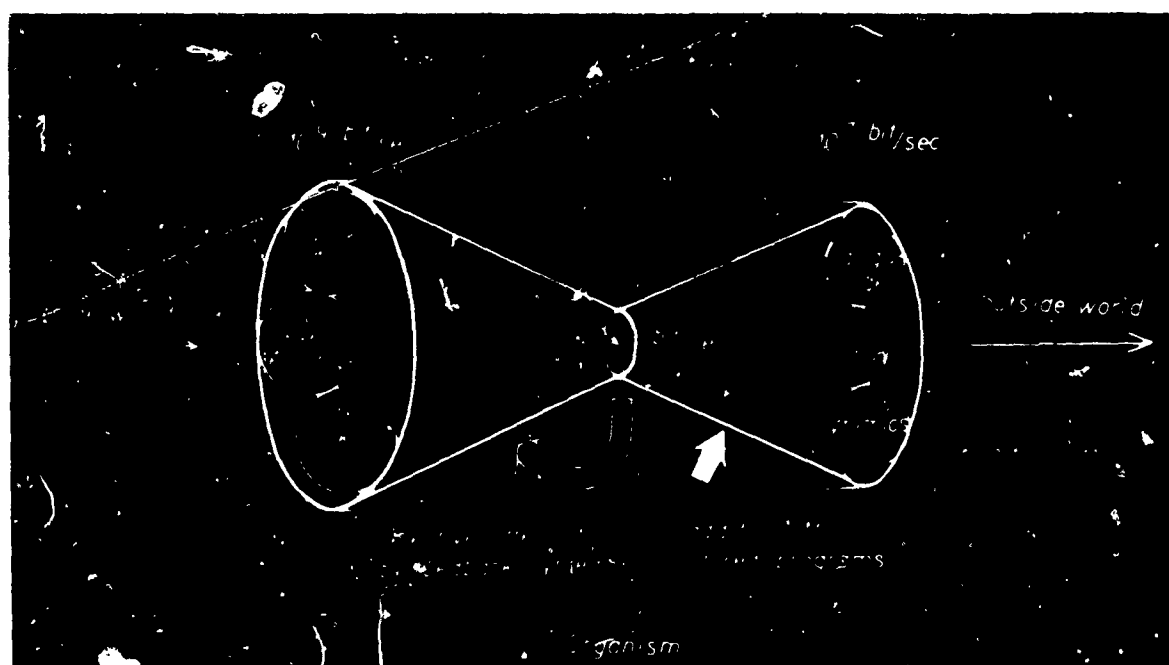


Fig.2 Prof.Keidel's representation of information processing by man



Fig.3 Photograph of a "naval battle" on a PDP-7 display unit

PAPER 16

EDUCATION

by

F. Liebesny

British Aluminium Co., U.K.

SUMMARY

The educational requirements for users and suppliers of scientific and technical information and the steps taken to provide professional education for various levels of attainment are discussed. Mention is made of specific efforts made in the U.K. to provide undergraduate and postgraduate training for users. The types of training courses available to suppliers - chartered librarians, library assistants and information scientists - are outlined.

EDUCATION

F. Liebesny

1. INTRODUCTION

The very magnitude of the problems connected with the information explosion is such that several approaches have to be made in the attempts towards overcoming them. Whether we use such almost meaningless figures as

- (a) that the world's output of scientific and technical articles is of the order of 1,000,000 per annum;
- (b) that the number of periodical titles in the disciplines of science and technology increases by two every day;
- (c) that about half of the world's literature in those self same disciplines is written in languages other than English (which for the purposes of this argument includes American English); or
- (d) that the periodical literature increases at a compound rate of approximately 6-7% per year,

we are still left with the almost irrefutable fact that all of us, users and suppliers alike, are gradually becoming submerged by this flood of paper. Although the argument of quality of the literature is being ignored in such discussions, it is obviously becoming increasingly difficult to cope with the output of the world's presses. This tendency - which is certainly not new - has led over the years towards more and more specialization, both in the user and supplier of the specialist literature. This has led on the one hand to the emergence of the information scientist, and on the other hand to, albeit, isolated endeavours to train the user in some of the more elementary forms of the techniques employed in coping with the literature.

The term 'information science' - though of comparatively recent origin - has acquired in this short space of time several shades of meaning. Therefore in order to define the way in which the term will be used in this paper, the activities embraced by the definition as given in the Articles of the Institute of Information Scientists should provide some guidance. These are:

- (i) abstracting, reviewing progress and other similar technical writing;
- (ii) translating scientific and technical writings;
- (iii) editing such writings as emerge from (1) and (2) above;
- (iv) indexing, subject classification and retrieval of scientific and technical information;
- (v) searching scientific and technical literature, preparing bibliographies, reports, etc.,
- (vi) providing scientific and technical information and tendering advice thereon;
- (vii) dissemination of information and liaison and field work for that purpose;
- (viii) research on problems in information work.

It is obvious from this recital that a professional information scientist should combine a considerable knowledge and skill for the proper execution of his work; normally this knowledge is obtained from courses of study while the skill should be derived from appropriate practical experience such as work in an information department. In order to attain a corporate membership in the Institute of Information Scientists it is thus necessary to provide evidence of both the required knowledge and skill; thus for Membership a candidate would normally be expected to possess a science degree and to have worked at least five years in an information department.

In training the user to enable him to deal competently with the documentation of his special subject field it would be unwise to aim at such a high degree of professionalism; firstly, a detailed training programme would turn the user into an information scientist and thus divert him from his own special field; secondly, it would be wasteful to impart knowledge and expertise of which a considerable amount would never be required by a specialist user of the literature since the full training of a documentalist involves matters relating to several disciplines; and thirdly, the time generally available for such training of the user is not sufficient for more than a somewhat superficial approach to the many problems of information storage and retrieval.

Therefore the few attempts that have been made to familiarize the scientist and technologist with proper means of using his subject literature require careful study to elicit their useful and successful features. The results of these courses are, however, not quite so easy to establish as most of them have only been conducted for a few years and it is thus still too early to quantitatively assess this criterion.

Before entering into a more detailed discussion of the courses available it must be stated - perhaps somewhat shamefacedly - that information on such educational ventures is not very easy to come by; it appears that the discipline of information science is not too well provided with means of keeping its practitioners informed on what is going on elsewhere. Although there are many periodicals and even abstract journals in the fields of librarianship, documentation, information science etc. their coverage on the educational front is not too extensive. It was for that reason that the Federation Internationale de Documentation (FID), the (British) Office for Scientific and Technical Information, Aslib and the Institute of Information Scientists organized in 1967 an International Conference on Education for Scientific Information Work. The proceedings of this conference were published in September 1967 by FID and its 33 papers tried to survey the international scene by focusing on the activities in the most important countries. That this attempt was not 100% successful can be deduced from the fact that there were no contributions from the USSR and that the Indian delegates did not attend the conference and thus take part in the discussions. However, these Conference Proceedings seem to constitute the most comprehensive review of the activities in educating users and suppliers of scientific information.

2. TRAINING OF USER

The user's training can be initiated at two levels: either before he graduates or afterwards.

Training at the undergraduate stage is frequently very difficult because most of the syllabi are so crowded that it requires considerable persuasion of the university authorities to devote any of that precious time to such peripheral subjects as documentation. Furthermore, there is still a great deal of that old belief that every scientist knows - or at least should know - the literature of his own subject. This fallacious attitude ignores completely any subsequent developments in the subject field and its literature or the use of modern techniques in dealing with it.

Training at the post-graduate stage, on the other hand, is likely to ensure that the recipient is in the proper frame of mind to accept the training as by then he is more mature and more aware of his needs with respect to documentation.

- (a) *Undergraduate training:* In the United Kingdom organized training of students is conducted at the Universities of Liverpool and Bradford¹ where the university authorities took the initiative, and at a series of six universities² (Edinburgh, London (University College, and Chelsea College of Science and Technology), Oxford, Warwick, and York) where the Office for Scientific and Technical Information is arranging for some 500 students to receive an information service and some instruction in its use.
- (b) *Postgraduate training:* the main centre for education in the use of literature at that level is undoubtedly the National Lending Library for Science and Technology, at Boston Spa, Yorks. where courses lasting about 10 working days have been run since 1965. A specimen timetable for such a course on the use of scientific literature is given in Appendix I.

A more recent development is now being supported by OSTI whereby retrospective searching of the medical literature by means of the MEDLARS technique (MEDical Literature Analysis and Retrieval System) will be made easier by providing facilities for consultation with specially trained liaison officers. Five such officers - at Newcastle, London, Edinburgh and two yet to be appointed - will help users in the formulation of search profiles and will advise on availability and capacity of this computer-tape index. Courses are also being held at the Hatfield College of Technology.

The above analysis is certainly not meant to convey the impression that there is only one type of user; indeed there are many different users who have, however, one common feature, viz. an innate reluctance - or even inability - to communicate their real needs for information to the potential supplier. It is therefore desirable that any training of the user should provide for some means of tuition in this delicate art of stating his actual requirements and not of hiding them behind surmises as to the possible location of the needed information.

3. TRAINING OF SUPPLIER

The education of the librarian, information scientist, documentalist, etc. is perhaps somewhat better organized than that of the user. Nevertheless, there are today so many different avenues of training towards producing a qualified information scientist that this plethora of facilities may create an impression of diffusion and even confusion. In order to reduce this seemingly unmanageable pile of information into some state of order it may be advisable to classify these data according to the type of supplier to be produced:

3.1 Librarian:

- (a) *Chartered:* organized full-time courses of two years duration are provided at several library schools which are housed within colleges of further education, polytechnics or universities.

Recent developments towards creating a degree in librarianship (as opposed to a post-graduate qualification as is provided at University College, London) under the aegis of the Council for National Academic Awards have led to the setting up of such a course at Newcastle to commence in the autumn of 1968. Other such courses are likely to commence with the following academic year.

- (b) *Assistant:* it was felt some time ago that a qualification of a lower level of competence than that of the chartered librarian should be provided for those people having to interrupt their professional life - especially married women - or those not wishing to attain the higher professional status. Towards this end, a Library Assistants Certificate has been proposed; this scheme will be administered by the City and Guilds Institute of London.

3.2 Information Scientist:

Since 1961 a post-graduate course has been run at the City University in London (formerly the Northampton College of Advanced Technology). This two-year course is run twice weekly (two hours each) to enable students to attain an entrance qualification for the corporate membership of the Institute of Information Scientists. A post-graduate one-year full-time course was started in 1963 which, since 1967, will lead to a M.Sc. degree.

Similar courses have been running at the University of Sheffield Postgraduate School of Librarianship and Information Science since 1964. At present this one-year course leads to a Diploma, but from October 1968 will lead either to a M.Sc. degree in Information Studies or in Librarianship.

Owing to the previously mentioned fact that about half of the world's literary output by scientists and technologists is written in languages other than English and that many of the readers of such writings are only too rarely qualified to comprehend those foreign languages, it is obviously important that the suppliers of information should possess some proficiency in handling foreign language material. Therefore many of these courses lay considerable stress on such ability and even include some training in languages in their syllabus. In the M.Sc. course at the City University one examination paper is devoted to testing the required level of proficiency.

While the foregoing survey has been directed largely to activities and developments in the United Kingdom it should not be thought that the rest of the world is standing still. Indeed, as R.T.Bottle³ has shown, similar schemes for training the user are in operation or being planned in many countries of Europe (to which his survey was confined) and in Australia⁴.

In the user-supplier dialogue which forms the theme of this symposium it is essential that each party should understand and appreciate the basic requirements of the other; this can be achieved, inter alia, by proper teaching and training in the elements of storage and retrieval of information for which usually too little time is available in the overcrowded timetables during the university courses. The few endeavours described above are hopeful omens that the need for a deeper understanding is being more widely realized.

REFERENCES

1. Bottle, R.T. In Proc. Intern. Conf. Educ. Sci. Inf. Work, 1966, pp.59-69.
2. Sommerfield, G.A. In Chemy in Britain, Vol.4, No.2, Feb.1968, pp.71-73.
3. Bottle, R.T., In J.chem. Educ., Vol.6, No.1, Feb.1966, pp-3-6.
4. Ward, J L (Ed) *Library Services for Chemists*. Melbourne, Australia, Royal Melbourne Technical College Press, 1960.

APPENDIX I

The use of scientific literature - A course for research students

TIMETABLE

Day	Time	
1.	2.00 p.m.	Tour/description/services/ of the National Lending Library for Science and Technology (in particular Reading Room and Staff Library).
2.	9.15 a.m.	Guides to published information <ul style="list-style-type: none"> i. Serials: Current awareness tools <ul style="list-style-type: none"> Abstracting journals Indexing journals Annual reviews Review serials
	10.30 a.m.	2. Tools for student's specified interest
3.	9.15 a.m.	3. Reports; Indexes
	10.30 a.m.	4. Books: Theses <ul style="list-style-type: none"> Annual Reports Yearbooks Monographs Technical dictionaries Language dictionaries Encyclopaedias Bibliographies
4.	11.15 a.m.	Language problems
5.	11.15 a.m.	Record keeping
6.	11.15 a.m.	Information bureaux
7.	11.15 a.m.	Keeping up with current literature
8.	11.15 a.m.	Library resources in the U.K.
9.	11.15 a.m.	Films. The National Lending Library for Science and Technology. National Library of Medicine, U.S.A.
10.	11.15 a.m.	Criticism and discussion of the course.

Time not devoted to lectures will be spent on literature searching.

(N.B. As can be seen from this schedule use is made of the few available films on this topic).

From: J.Doc., Vol.22, No.1, March 1966, pp.22-32

DISCUSSION

A.H.Holloway: In an instructional tour for new entrants to the British Royal Naval Scientific Service, it has been found that over half of these graduates have never had any instruction in the use of the literature and that after about six months' working experience about one third of them have never been into their establishment libraries. At the request of these new entrants short courses are being arranged to demonstrate the information facilities available and how to make the best use of them.

R.R.Dexter: In the U.S.A., the Institute of Aerospace Sciences has made special efforts to make the engineer in industry aware of information and documentation facilities. The approach is to send an engineer with experience in information work and skilled at putting his ideas across into a firm, where he investigates the documentation activities and suggests improvements to and better ways of using the system.

F.Liebesny: This is a very interesting approach. In any activity of this kind, full support from management is essential.

SUMMING UP OF THE TIP-SYMPOSIUM

R. Bree

I think it is an almost impossible task to summarize the impressions gained from the presentation of sixteen well prepared and extremely stimulating papers as they covered a wide range of subjects. Besides, in some cases they touched topics outside my own experience and I am therefore not entitled to give any judgements. This situation has been made even worse as circumstances made me a victim of a faulty information transfer: when asked by a rather blurred long-distance call to take over the chair for the last session - at least this is what I understood - I agreed, only to find out later on that I had quietly committed myself to trying to produce this summary. In other words: something looking initially like a privilege amounted in fact to a sentence to hard labour!

That is why I am asking to be forgiven for changing what was called "summing-up" into the presentation of a rather incomplete, thoroughly subjective and probably biased account of my personal impressions.

In so doing I am wearing three different hats.

First, being an engineer and at the same time director of the nuclear documentation and information centre of the European Communities (Euratom/CID) I am certainly a supplier of information, responsible for the development of the largest European venture in mechanization: the Euratom Nuclear Documentation System, which is giving access to a store of over 750,000 items of document data and which is in operation for retrospective searches and for SDI.

Secondly, all my life I have been a user of technical information.

Thirdly, as I write a paper every once in a while, I must confess to being also an originator of such information.

Result: the conflicting views and feelings which are involved in wearing of three different hats are struggling at this moment within me.

To start with, the title of the symposium. I am afraid this was more or less a misnomer. I simply missed any true dialogue between suppliers and users. Didn't we in fact hear yet another extensive multifaceted monologue of those who claim to be suppliers of those often extremely perishable goods which it is customary to bundle under the term "scientific and technical information"? I really wonder how the members of the co-sponsoring Avionics Panel feel about this!

On the other hand, it is not surprising that there was no real dialogue with "the user". The user does not exist; what exists is a rather inarticulate mass of users, each of them full of individual expectations, insofar as they expect anything at all.

The first paper gave me reason for some embarrassment: it retold the history of all the brave attempts to do some reasonable spadework to improve the conditions on which real progress in information handling does depend. These attempts were at that time mainly inspired by suggestions from AGARD/TIP but they seem to have been shelved and forgotten.

In the meantime, their revival is being undertaken by bodies like OECD and UNESCO. This is not what is embarrassing me but the little response received from governmental management and administration in following up the recommendations we presented about 8 years ago. One would assume that for decision-making, that outstanding governmental privilege, everybody involved would crave for pertinent information on which to base decisions. It might well be, however, that so-called political decisions have to follow a different pattern.

The next observation - and this might be rather biased - is how differently the scientist or the engineer tends to act, depending upon which position he assumes momentarily: that of the user, or that of the originator of information. Being rather demanding and not easily satisfied as a user, he seems inclined, as an originator, to be forgetful of all the rules of good behaviour and of all the necessities for making sure that his bit of information becomes easily retrievable later on (by giving it precise bibliographical references, exact descriptive data for cataloguing and a decent abstract of informative value).

What has struck me generally was how much attention has been devoted during this symposium to all the problems of the hardware involved and how to handle it aptly, and how little attention to all the questions of quality control at the originating level. It goes without saying that in this respect the colleagues handling information related to defence are hit hardest: disseminating information and respecting security regulations are definitely incompatible and often even contradictory. The high amount of public money spent in this field in combination with the alleged necessity for secrecy seem to favour the research report as a means for the presentation of the results achieved. If and when the stamp "secret" is added, the number of readers is anyway reduced and so is the chance that considerations concerning the quality of the contents come into the picture. I think that the real chance and hope for reducing the amount of garbage lie in publishing a maximum of results through journals of high standing, the publishers of which maintain commendable quality standards for the articles they accept for publication.

★ ★ ★ ★ ★ ★ ★

Another group of problems concerns the intrinsic value of information - mechanization in itself. We found amazingly optimistic views on the future development potential of the equipment, which sometimes seemed to be extrapolated from far too modest statistical data to suggest reliability of the conclusions. From my own experience I know that many a problem shows its real and overwhelming dimensions only under real life conditions and not when appraised on too small samples. Parts of such optimistic statements seemed to be slightly influenced by the desire to introduce machinery with technically interesting features in this up-and-coming field of information handling, without worrying too much about the economics involved. Opposition against these optimistic views on the usefulness of sophisticated machinery for information handling was actually voiced in cases where manual methods proved sufficiently effective. This was illustrated by the very impressive achievement of the Netherlands centre which stated that 250 requests per day constitute the average workload of its manually operated system.

But I think that even this very admirable feat cannot stop the trend toward the introduction of mechanical methods. The results obtained by the Euratom System support this statement, as does the firm intention of some editors, like those of Chemical Abstracts, to mechanize their operations and use to a full extent all the gains in speed, accuracy and quantity which are possible with an extremely careful data-preparation. It is remarkable, by the way, that for the moment all operational mechanical systems depend upon the skilful combination of scientific and technical staff for handling the phases of literature subject control, vocabulary control and systems development. Part of this staff is indispensable for the translation of the customers' natural language questions into machine language, for screening the results provided by the machine, etc. I am convinced that this will stay with us for quite a while, even though higher degrees of automation in the introduction of data might prove feasible. I suppose that for the often advertized dialogue between user and machine, such knowledgeable interpreters are not only essential but their employment seems to be - at least for the time being - the most effective and the most economic answer to the problem of ensuring satisfactory service to the customer without having to train large numbers of potential users.

Introducing machinery in information handling unavoidably means introducing centralization of operations on the one hand and partition of work on the other. This is new for this field and needs a rethinking of relations and working style. The capacity of computer-based information systems to handle numerous individual requests and the still high cost of computer time would lead to rather high operating costs, if untrained users were allowed direct access to the computer. The customer needs an interpreter, i.e. a member of the information centre with the needed amount of knowledge in both subject matter and retrieval processes, to obtain fast and inexpensive answers.

I feel that it is unrealistic to-day to believe that large numbers of potential users can be properly trained to draw full benefit from mechanized systems by posing their questions directly to the machines. One cannot hope that the rather small number of specialists who have developed and who now operate existing systems could possibly provide user training on a large scale; failing this thorough training, the user is left to his own devices such as haphazard methods of trial and error which not only cost machine time but may also shy him away if the results are not up to his expectations; it is indeed a common human habit to blame others, including machines, rather than blame oneself.

It is precisely because I am convinced that machines can lead to a much better use of existing information, whatever its volume, that I feel that much care should be devoted to make users familiar with these new methods, thereby ensuring them access to the documents they need.

The discussion on the economics of information handling was most interesting. I am afraid, however, that the useful starting point for assessing the economic impact of the use - or possibly, of the non-use - of information has still to be looked for. To my understanding, creating a successful system of storage and retrieval of information within a given field of interest is only tackling one half of the problem, even though this system might yield excellent results. Creating such a system means only creating a potential. The real value of the potential does not only depend on the supplier of the system and not even on its quality; it depends upon what the user is drawing from the system, first in form of access to available information and secondly, by the user's own and irreplaceable act of evaluating the accessible information for his own problem and purpose.

The problem of how to measure at all such an impact in its various consequences, positive or negative (i.e. taking advantage of given information for acting in a way one would not have acted otherwise or for avoiding an obvious mistake) seems to be almost impossible, especially if one wants exact and convincing results.

Undeniably, the high expenses of creating machine systems constitute a formidable obstacle as they must be wrung out of parliaments which are slightly scared by the complexity of the problems involved. I feel therefore that much effort must be concentrated on all factors which influence the economics of the development and even more the operation of such systems. The potential such systems are offering is of identical interest to any industrial country in the world. These countries are also the largest originators of technical and scientific information. By cooperation they can share the burden of input in such systems, and this is indeed a considerable burden because so much intellectual work is involved. Sharing the burden means, at the same time, individually enjoying all the advantages of using a common and centrally processed input. Another means of improving the overall economics of such systems is doubtlessly the formulation and strict observation of a set of standards for presentation, codes, data-display, etc. and, not less important, the acceptance of minimum quality standards for abstracting.

Still another problem would be solved if an acceptable way could be found to meet expenses of operating systems by the establishment of fair charges for their use. I came here wondering whether or not this ticklish point would be touched by one of the papers read or during the discussions. Alas - I am taking leave empty-handed as far as this problem is concerned but I am open to future reasoning on this topic.

A higher degree of input automation is recommended for the sake of economy, too. Rather complicated measures for vocabulary control and less expensive machine memories would be indispensable. But even with this, a highly flexible character reading machinery would still be a necessity.

The mentioned data banks seem to offer much promise on the condition that access to documents containing valuable data is guaranteed and that, furthermore, a very considerable scientific effort is performed: the data as contained in the documents are of value, in the majority of cases, only when they are related to common standard bases and are therefore comparable. The needed standards are not yet agreed upon to my knowledge. The extremely high-life-expectancy of the information established by such transformed and comparable data should justify the considerable effort which has to be made prior to their establishment and storage.

* * * * *

The old topic of "information" versus "documentation access" showed up as usual. It seems to me that information really becomes information solely through the evaluation and perception of a message by the user himself. This process is not facilitated but hampered by any form of predigestion whatsoever, even if this predigestion is done for any group of users of seemingly identical interest. The value of one and the same document can be much at variance depending on the problem it is supposed to solve. This seems to lead to the conclusion that it is wiser to concentrate efforts on providing fast and dependable access to any group of documents corresponding to a stated subject-interest. Wouldn't it be wonderful if we could claim that this first and indispensable step had been satisfactorily achieved?

In several places the use of free language has been advocated during this symposium. Is it really conceivable that this can be brought about in multilingual systems? I am afraid that the corresponding needs for very large machine memories would exceed all available resources, at least financial ones.

* * * * *

Coming now to the summing up in the summing up. I can only state that the problem remains as complex as before; however, for its solution, as much can be done by improving basic methods of recording the information, applying useful standards, carrying out intensive training in information handling and creating information centres staffed with subject expert teams as by pressing on in the direction of more and more automation.

I am afraid that the striking difference between the tremendous speed of machine-development on the one hand and the much more modest human capacity for adaptation to the new methods offered, will be one of our gravest problems when introducing machine methods to a "clientèle" consisting of individuals. Speeding up this adaptation must therefore be an integral part of our effort.

Certainly, we must never prevent research and inventiveness from assisting us in useful forms of information handling. But we must be very careful in applying the new methods and very thorough in testing their impact on the customers. The financial limitations for system building will stay with us, at least here in Europe, for a very long time; to increase our resources, the indispensable effort of convincing politicians and administrators of the usefulness of using information might well add to our burden.

Looking back at all the stimulating papers and discussions, one thing seems clear to me regarding the future rôle of scientific information, especially in our highly industrialized countries: whether information has a measurable economic value or not, it is tempting to take over the famous definition somebody gave of the term "tact":

"When you don't have it, you must see how you can do without!"

But I am fully convinced that our countries cannot do without full access to the world potential of existing scientific and technical information.

VOTE OF THANKS BY H.F. VESSEY, CHAIRMAN T.I.P.

Mr. Chairman, ladies and gentlemen, it is my privilege and pleasure on behalf of AGARD to express our thanks to all who have contributed in making this symposium a success.

Director Finn Lied our Chairman, who opened the sessions regrets that he cannot be present but is very conscious of the assistance that has been provided by the German Ministry of Defence and in particular by Dr. Beneke, the German National Delegate to AGARD. In this building the DGLR has done a wonderful piece of organisation in providing magnificent accommodation and facilities and in particular I must congratulate Mr. Steckel on the smooth way in which the flow of nearly 200 persons has been unimpeded. When we discussed the meeting arrangements with Mr. Steckel, frankly I did not believe that coffee breaks could be held to 20 minutes; his confidence in his staff has been fully justified. Dr. Rautenberg too has helped considerably in the arrangements and I wish to thank him and his staff. I am sure that you will agree that the interpreters, sound technicians and projectionists have done extremely well.

Finally, I want to thank the authors, session Chairmen and last but not least, you, the audience. I have been a little disappointed at the reaction from the "users", you must be more satisfied with your documentary services than I expected. There is still time, however, if you wish to comment, or ask further questions, on papers you have heard. Take a question paper, fill it in and send it to AGARD in Paris.

I am sure there are many others I should thank for their assistance, time is short, however, and I must apologize for omitting them and say that the gratitude is no less genuine.

Most of us have had time only in the evenings to see the beauties of Munich but I certainly have been impressed by the Rathaus and by the friendliness of the ordinary Munchener.

I am myself too close to the subject and too much involved to assess the results of the symposium objectively but I must say that I am very satisfied.

Once again, our thanks to all our German friends and "Auf wiedersehen".

NAME INDEX

- Altman, J.W., 63, 69, 72, 76
 Angell, T., 86
 Arenstorf, R.F., 94
 Ashburn, E.B., 166
- Baker, F.B., 86
 Bell, C.A., 20
 Boldovici, J.A., 69, 72
 Borko, H., 83, 85, 86
 Bosman, D., 10, 76, 149, 167
 Bottle, R.T., 184
 Bourne, C.P., 86
 Brackett, J., 156, 166
 Brear, R., 20, 39, 88, 187
 Brown, S.C., 162, 166
- Cheek, T.B., 166
 Christensen, W.C., 20, 123, 131, 132
 Christman, P.A., 166
 Cleverdon, C.W., 85
 Corbato, F.J., 166
- Day, M.S., 61, 133, 149
 Dexter, R.R., 186
 Dubon, R.J., 21, 39, 61, 149
- Edwards, J.S., 84, 85
 Einsele, T., 167
 Engellen, G., 94
- Fano, R.M., 165
 Feidelman, L.A., 10, 49, 58, 61
 Fischer, G.L. Jr., 58
- Gabelman, I., 167
 Gabrini, P., 86
 Gandy, R.W.G., 131
 Gardin, J.C., 81, 85
 Garvey, W.D., 72
 Genrich, H.J., 94
 Gerhards, L., 94
 Greenberger, M., 156, 166
 Griffith, B.C., 72
- Hale, J.F., 69, 72
 Hamblar, K.K., 58
 Handler, W., 169, 176
 Hangsted, F., 131
 Harman, R.J., 166
 Herner, S., 72
 Holloway, A.H., 39, 88, 131, 186
- Isotta, N.E.C., 111, 115, 116, 122
- Jones, M.M., 156, 166
- Kanal, L.N., 58
 Kaplow, R., 156, 166
 Katz, L.N., 58
 Keidel, Prof., 173
 Keonjian, E., 39, 131
 Kerr-Waller, R.D., 10, 39, 131
 Kersta, L.B., 58
 Kessler, M.M., 166
 Kruckeberg, F., 89, 94, 97
- Lapeysen, E., 39
 Lefkovitz, D., 86
 Lengyel, B.A., 166
 Lesk, M.E., 86
 Lavery, F., 71
- Levinthal, C., 156, 166
 Licklider, J.C.R., 76, 115, 151, 167
 Liebesny, F., 179, 186
 Lindenberg, W., 94
 Little, J.P., 176
 Lustig, -, 97
- McGill, D.W., 72
 Menzel, H., 72
 Mey, J., 94
 Molzberger, P., 61, 176
 Montgomery, E.B., 41
 Moreh, J., 120
 Morris, J.H. Jr., 156, 166
 Moser, R., 149
 Moses, J., 156, 166
 Munger, S.J., 72
- Needham, R.M., 86
 Ness, D.N., 156, 166
- Ossorio, P.G., 67, 72
 Overhage, C.F.J., 165, 166
- Payne, D., 69, 72
 Paddock, B., 58
 Pollack, D.K., 58
 Price, D.J. de Solla, 65, 72
 Prywes, N.S., 77, 86, 88, 97
- Rath, G.J., 72
 Resnick, A., 72
 Rosenfield, A., 58
- Sager, N., 85
 Salton, G., 85, 86
 Saltyer, J.A., 166
 Savage, T.R., 72
 Sayers, W.C.B., 85
 Schjetne, K.G., 10
 Schnelle, H., 94
 Schrader, R., 110
 Schuler, S.C., 10, 110
 Schüller, J.A., 99, 110
 Schweisthal, K.G., 94
 Skoumal, S., 61, 167, 176
 Sommerfield, G.A., 184
 Scollar, I., 94
 Speiss, W., 10
 Stark, R., 167
 Stevens, M.E., 58, 82, 85
 Stolk, H.A., 110, 115
 Stone, P.S., 81, 85
 Stotz, R.H., 166
 Strong, S., 156, 166
 Swanson, D.R., 86
- Tanyolac, N.N., 10
 Toma, P., 93, 94
- Unger, S.H., 56
- Vernimb, C.D., 110, 116, 149
 Vessey, H.F., 10, 11, 20, 61, 88, 122
 Vickery, B.C., 72
- Ward, J.E., 166
 Ward, J.L., 184
 Wenke, K., 94
 Williams, J.H., 85
 Wolfe, J.N., 117, 122
 Wright, R.C., 10, 39

Entries in italics indicate authorship of a chapter

SUBJECT INDEX

- Abstracts versus full test for retrieval 69
- Abstract journals, use of at TDCK 102
- Access, bibliographic 163
 - controlled 158
 - direct in ESRO/ELDO 115
 - physical retrieval 163
- Accession lists 135
- Accumulated software 157
- ADMINS program 156
- Alexandria, library at 4
- Algorithmic processing 80,
- Alphanumeric readers *see* Readers
- Analysis, multivariable statistical 67
 - syntactic and semantic 80
 - and synthesis 70
 - of text 69
- Announcement cards (SDI) 137, 140
- APEX System 154
- Application, Science of 43, 45-46
- Artificial intelligence 83
- Association terms 80
- Augmented catalogue 163
- Automata 157, 175
- Automatic classification 82
- Automatic indexing 82

- Bar code optical readers 53
- Bibliographic, access 163
 - file 66
 - review 68
- BOLD system 83
- Boolean logic 136
- Browsing, in computer system 115
- BULL - General Electric font 52, 59
- Business activities in documentation,
 - use of computer for 113

- Card indexes (subject), used at TDCK 103
- Catalogues,
 - augmented 163
 - preparation by computer 162
- Chain reaction in generation of
 - information 46
- Character recognition 52
 - application 55
 - future trends 56
- Charging for information services 129
- Circular Thesaurus *see* TDCK Circular Thesaurus
- Citation 79, 84
- Citation indexing 17, 66
- City University, London, training course
 - in information science 184
- Classification Automatic 82
 - (military security), effect on communication 5
 - subject 67
- COGO (Co-ordinate Geometry) Language 155
- Communication, Science of 43, 44, 46
- Compact System *see* TDCK Compact System
- Compagnie des Machines BULL - General Electric
 - font *see* BULL - General Electric
- Computers, GEC 645 154
 - IBM 360 154
 - IBM 360/40 154
 - IBM 360/67 154
 - IBM 1130 154
 - IBM 7094 154
 - PDP-6 154
 - PDP-7 176
 - TX-2 154
- Concordance generation 81
- Consoles 154
- Controlled access 158
- Co-ordinate indexing 83,
- Costs, information services in NATO countries 119
 - library services in UK 119
 - US Military Specification programme 127
- CP System 154
- Critical Reviews, Value of 7
- CTSS (Compatible Time Sharing System) 154
- Current awareness 68, 135
- Cybernetics 45, 46

- Data, engineering, handling of 127
 - technical, definition of 125
- Defense Documentation Centre, U.S.A. 128-129
- Defense Documentation Centres of NATO 13
- Department of Defense, U.S.A., information
 - services 128-129
- Descriptors 67, 105
- Dewey Decimal Classification 81
- Direct access service, ESRO/ELDO 115
- Directory generation, a posteriori 81
 - a priori 81
- Distance (between documents) 82
- Documentation centres, functions 13
- Displays 154, 172
- DYNAMO Language 155

- E-13B standard font 52, 59
- Editing, using interactive computer 155
- Education for Scientific Information Work,
 - International Conference 1967 182
- Education of documentation users and suppliers
 - see* Training
- Efficiency of retrieval, definition 65
- ELDO *see* ESRO/ELDO
- ESRO/ELDO Space Documentation Service 114

- European Launcher Development Organisation
 - see ESRO/ELDO
- European Space Research Organisation,
 - see ESRO/ELDO
- Evaluation of text 70
- Exchange of reports 102
- Extra ts 69,71

- Fact retrieval (definite answers) 163
- Feedback, from SDI users 136,138,140
- Flying spot scanner 52
- Fonts, standard, character recognition,
 - BULL - General Electric 52,59
 - E-13B 52,59
 - ISO-B 53,60
 - USASI 53,60
- Formulation of query 84
- Forschungsgruppe Linguistik und
 - maschinelle Sprachubersetzung see
 - LIMAS System
- FORTTRAN 155

- Games theory 91
- Graph theory 93
- Graphic processing 157
- Group profile 139
- Group theory 91

- Handwriting, recognition of 56,60
- History of documentation 3

- Identifiers 163
- Incremental simulation 156
- Index terms 136
- Indexing 126,128
- Information 44,
 - definition 65
 - dynamics 46,
 - presentation in technical writing 71
 - processing 153
 - retrieval 30,158
 - Science of 43,44,46
 - sources, personal 127
 - sources, used at TDCK 102-103
- Information Analysis Centres 15,17,128
- Information Science, definition 181
- Information Scientist, duties 128,181
 - training 184
- Input to computers 32,114
- Institute of Information Scientists 181
- Interactive retrieval 158
- Interactive program 155
- Interactive query 84
- INTREX Project (Information Transfer
 - Experiments) 163
- "Invisible colleges" 66,135
- ISO-B font 53,60
- IVERSON Language 175

- Journal categorization 135

- Keyword retrieval 24,83
- KWIC index 16

- Language (meta) 153
 - processing 79,81
- Languages, programming
 - COGO 155
 - DYNAMO 155
 - FORTTRAN 155
 - IVERSON 175
 - LISP 155
 - MAD 155
 - STRESS 155
- Learning machines 53
- Lectriever, used in TDCK 103
- Librarians, training of 183
- Libraries, forecasting demand for 119
 - function as information suppliers 15
- Lightpen 172
- LIMAS system 92
- LISP (List Programming Language) 155

- MAC Project (Machine Aided Cognition) 153
- Machine translation 92-93
- MAD Language (Michigan Algorithm Decoder) 155
- Magnetic reader 54
- Man-computer interaction 79,157
- Management, as information user 125
- MAP Program (Mathematical Assistance Program) 156
- Mark-sense optical readers 53
- Matrix matching 53
- Mechanisation of documentation, need for 113
- Medical diagnosis, application of pattern
 - recognition 54
- MEDLARS liaison officers 183
- Meta-language 153
- Microfiche 18,164
- Microforms 17,
- Modelling 156
- Molecular structure, programs to display 156
- Monte Cassino, library at 4
- Multi-access 80,157
- MULTICS Program 154
- Multi-dimensional scaling 69
- MULTILIST System 83
- Multi-variate analysis in classifying documents 67

- Natural language processing 81
- Nervous system, analogies with computer
 - system 173
- Netherlands Armed Forces Scientific and
 - Technical Documentation and Information Centre
 - see TDCK
- Network integration 163
- NEWIP Project 159
- Notification listing in SDI 138,140

OCR see Optical Character Readers

On-line, display 83

retrieval 83, 157

OPS 156

Optical Character Readers 114

Orthogonality 67

OSTI-OECD study of economics of
information systems 119

Pattern recognition 53, 93

applications 54

future trends 56

Photograph-interpretation, application
of pattern recognition 54

Precision ratio 84

Preprocessing 79

Processing, graphic 157

language 79, 81

Profile, group 139

interest 136

standard 139

topical 139

Programming languages, see Languages,
programming

Programs

ADMINS 156

APEX 154

CTSS 154

CP 154

MAP 156

MULTICS 154

OPS 156

SORTER 155

TEACH 156

TYPSET 154

Program correction using computer 155

Project

INTREX 163

MAC 153

NEWTIP 159

SMART 84

SYSTRAN 93

TIP 158

Query 83

interactive reformulation 84

Question asking 156

Readers

bar code 53

curve tracing 53

mark sense 53

IBM 1287 56, 60

costs 54, 57

Reading, mechanical 18

Recall 66

Recall ratio 84, 139

Recognition, character 52

pattern 53, 93

Recognition Unit 52

RECON (Remote console) 142

Referral Centres 16, 17

Reformulation of Query, interactive 84

Relevance ratio 139,

Retrieval 66-68, 83, 158

effectiveness 84

SCAN (Selected Current Aerospace Notices) 140

Scanners, see Readers

Science Committee of NATO 6

Screening of documents 66, 68

SDD (Selective Dissemination of Documents) 141

SDI (Selective Dissemination of Information)

17, 69, 71, 135-142

Search Input 141

Semantic analysis 81

Sheffield Univ., course in information

science 184

Simulation (incremental) 156

SMART Project 84

Soft copy 164

Software 156

SORTER Program 155

Specialised Information Centres see

Information Analysis Centres

Specifications, Military, issued by Dept. of
Defence U.S.A. 127

Staffing at TDCK 101

STAR bulletin (NASA), input by ESRO/ELDO 114

Statistical techniques 82

Storage 80

STRESS Language 155

Stroke analysis 53

Suffix editing 81

Symbolic analogue 67

Symbolic integration 82

Syntactic Analysis 80

SYSTRAN Project (Russian-German translation) 93

System analysis 46

Systems

APEX 154

FOLD 83

CP 154

LIMAS 92

MULTILIST 83

Systems, Science of 43, 45, 46

Tape exchange 18

TDCK

functions 101

organisation 101, 107

system used for retrieval 104-106

Circular Thesaurus, 105

Compact System 104

TEACH Program 156

Technical Information Panel (TIP) of AGARD 6

Text, definition 65

Textual information 65, 71

analysis 69

Textual Unit 70,71	Transport Unit 52
Thesauri 26,79,81	TYPSET program 154
Time Sharing 18,158	
TIP (Technical Information Program) 158	USASI font 53,60
TIP (Technical Information Panel of AGARD) <i>see</i> Technical Information Panel	User feedback (SDI) 136,138,140
Topical profile 139	User needs 121,126
Training, of suppliers 19,183-184	
of users 8,19,182-183	Voiceprint 54
Transfer of knowledge, definition of 65	
Transformations 156	
Translation, by computer 92-93	Word stems 81,83

This AGARD Publication can be obtained from:

Clearinghouse for Federal Scientific and
Technical Information (CFSTI),
Springfield, Virginia 22151, USA.

The unit price for a hard copy is \$3.00. The unit price for a microfiche copy is \$0.65. Payment should be by check or money order and must accompany the order. Remittances from foreign countries should be made by international money order or draft on an American bank, payable to CFSTI. The request for a document should include the AGARD Report Number, title, author and publication date.

If you are interested in obtaining other AGARD publications consult either the -

Scientific and Technical Aerospace Reports
(STAR) published by the Office of Technology
Utilization, National Aeronautics and Space
Administration, United States Government,

or the -

United States Government Research and Develop-
ment Reports Index (USGRDRI) published by
the Clearinghouse for Federal Scientific and
Technical Information, US Department of
Commerce.

National Aeronautics and Space Administration
WASHINGTON, D. C. 20546

OFFICIAL BUSINESS

POSTAGE AND FEES PAID
NATIONAL AERONAUTICS AND
SPACE ADMINISTRATION

05U 001 11 66 3DS 69304 02672
DEFENSE DOCUMENTATION CENTER
SCIENTIFIC AND TECHNICAL INFORMATION
CAMERON STATION, BLDG. 5
ALEXANDRIA, VIRGINIA 22314

ATT DDC RESEARCH LIBRARY



Printed by Technical Editing and Reproduction Ltd
Harford House, 7-9 Charlotte St. London. W.1.